ulm university universität ulm

Ulm University | 89069 Ulm | Germany

Faculty of Engineering, Computer Science and Psychology
Institute of Databases and Information Systems

Evaluating a Configurator Application for Modeling Data Collection Instruments: An Experimental Study

Diploma Thesis at Ulm University

Submitted by:
Dominic Gebhardt
dominic.gebhardt@uni-ulm.de

Reviewer:
Prof. Dr. Manfred Reichert
Dr. Rüdiger Pryss

Supervisor:
Johannes Schobel

2016

Version November 11, 2016

Satz: PDF-LaTeX 2$_\varepsilon$

Abstract

Software applications are a crucial part of our daily life and become more and more indispensable. Over the years, software itself has gone through a rapid development. Therefore, not only the complexity has increased, but also the quality requirements of customers. One important quality aspect that have to be fulfilled in order to guarantee acceptance of the software application is its usability. However, ensuring the latter is by no means easy. Thus, developers are obliged to test their software product during the development phase in order to gather respective informations. Furthermore, testing allows reacting immediately and adapting the software accordingly if needed. Experiments are common methods for testing and, therefore, for obtaining important data as scientific foundation for further analysis.

The thesis at hand evaluates a newly developed configurator application for modeling data collection instruments regarding its usability. Therefore, a controlled experiment is set up in order to investigate how much effort end-users need to properly handle the application. In particular, the overall understanding of the (modeling) concept with respect to the complexity of the configurator is evaluated. In this context, the software is used in order to model specific data collection instruments. More precisely, participants in the experiment are asked to process tasks of different complexity using the configurator application. Thereby, the focus is on the total time needed to solve the tasks as well as the number of errors in the resulting questionnaire models. Considering the results obtained assumptions regarding the intuitiveness of the (modeling) concept can be made.

Acknowledgments

First, I would like to thank my supervisor Johannes Schobel for supporting me. His assistance, feedback as well as guidance was priceless.

Furthermore, I would like to thank my first reviewer Prof. Dr. Manfred Reichert as well as my second reviewer Dr. Rüdiger Pryss.

I would like to acknowledge all participants for their contribution to this work.

A very special thanks goes to my brother Timo Gebhardt for providing feedback as well as for his invaluable moral support.

Last, but not least, I would like to thank my parents for their patience. Their inconceivable encouragement and support made this thesis possible.

Contents

1

Introduction

Software applications have become ubiquitous in our daily life. Whether within social or business context they have great influences of our behaviors. Almost every device integrates more or less complex software in order to simplify our daily work. This could range from simple devices like pocket calculators over more complex systems of, perhaps, heating systems to highly complex control systems for airplanes. Therefore, software applications are indispensable. Over the years, software has gone through a rapid development and its development will be just as quickly in the future. Thus, the software engineering process itself has become more and more important not only with respect to an increasing software complexity but also regarding the software requirements. Especially, when the objective is to care about the safety of human lives (i.e., air traffic or medical domain) robustness as well as reliability are crucial. However, less sensitive domains (i.e., data collection in health care) also have high quality demands on nowadays software applications. Therefore, two sides have to be considered. On the one hand the end-users who have to deal with the final product must accept the software. On the other hand the developers who want their software to be sold and, thus, to be accepted. But how can developers determine if their software product will be accepted by end-users during the development phase? In this context, software evaluation becomes a crucial part. However, the software engineering process itself is commonly based on observations but not on scientific research. In [1] the need for a scientific basis for software engineering is stressed. The latter is claimed as a laboratory science and, therefore, empirical research methods in the field of software engineering should be applied. This allows for efficiently collecting informations as a scientific foundation for software quality improvements. Considering the quality aspects,

one important factor (besides others described in [2, 3]) when developing software applications is the usability of the application. Developers have to guarantee that their product will be user-friendly regarding the complexity of its application. Otherwise, the software application developed will most likely be not accepted. In the context of this thesis, an empirical research method is used to evaluate a configurator application for modeling data collection instruments regarding its usability.

1.1 Contribution

Considering the usability of a software application, the developers have to ensure that their product meets the high requirements of customers. In order to collect respective informations it is recommended to test the software during the development phase. This allows reacting on discrepancies and, therefore, adapting the software if needed accordingly. [4] introduces different methods allowing to gain informations about the object of research (i.e., software application). Experiments, inter alia, are common methods in order to gather data to provide a scientific foundation for further analysis. An advantage when conducting experiments is that the situation is completely controllable. This allows for manipulating the behavior systematically if needed [4].
In the context of this thesis a new software application developed at the *Institute of Databases and Information Systems at Ulm University* shall be evaluated. This graphical application allows end-users to model data collection instruments. Since data collection becomes necessary in many domains (i.e., health care) the need for simplifying the procedure itself as well as for evaluating the obtained data has increased. Therefore, the newly developed software application provides the possibility to model questionnaires in order to support the collection of respective data. Since the use of the software should be independent from different background knowledge (i.e., information technology) it is recommended that the concept is intuitive. The application should not only be used by experts. In this context, an experimental study is conducted in order to evaluate the usability of the mentioned software application.
The goal of the experiment is to investigate whether or not end-users are able to properly handle the introduced software application. In particular, the overall understanding of the

(modeling) concept with respect to the complexity of the application shall be evaluated. For this intention, different tasks are developed, which the subjects (i.e., participants in an experiment) have to process only using the provided application. The subjects are divided into two groups. One group has less (or no) prior experience in process modeling while the other group already has experience. Considering the differences in total time needed to solve the tasks as well as the number of errors made in the resulting questionnaire models the intuitiveness of the (modeling) concept can be evaluated.

1.2 Structure of the Thesis

This thesis is divided into seven chapters (cf. Figure 1.1). Chapter 1 motivates the thesis and discusses the contribution. Fundamentals are described in Chapter 2, while Chapter 3 presents the experiment definition and its planning. The experiment preparation, execution and data validation is described in Chapter 4. Chapter 5 presents the analysis of the results including descriptive statistics, data set reduction, hypothesis testing as well as an interpretation and discussion of the obtained results. Finally, Chapter 6 discusses related work, while Chapter 7 summarizes the thesis.

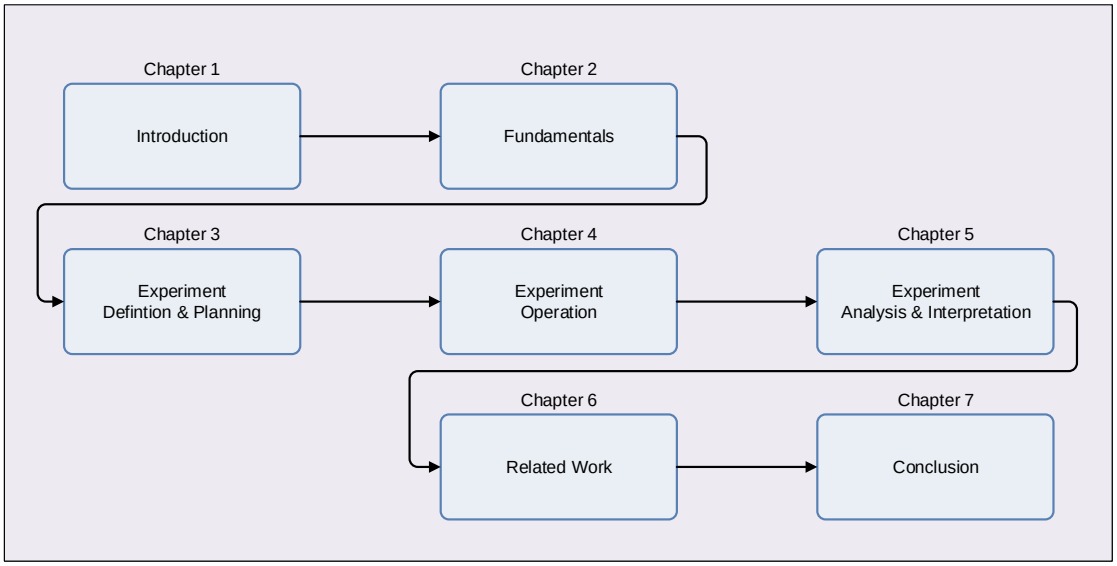


Figure 1.1: Structure of the thesis

2

Fundamentals

This chapter introduces fundamentals which are important for understanding this thesis. Section 2.1 introduces the *QuestionSys* approach, while Section 2.2 describes the *Configurator* component in more detail.

2.1 The QuestionSys Approach

Nowadays, smart mobile devices and its applications become an integral part of our daily life. This technology offers many possibilities for different life domains. Especially in domains processing a massive amount of data, mobile applications may support stakeholders and facilitate their work. Inter alia, it offers an alternative to collect patient data more effective in medical domains (e.g., psychology and healthcare), where data collection is traditional paper-based. Additionally, this manual data acquisition is a time-consuming and costly process. Therefore, domain experts demand new concepts to speed up the process of data collection and make it more effective. However, not every domain has the possibility to spent plenty of money for IT-Experts nor has the know-how to develop complex applications by themselves [5, 6]. Hence, the *QuestionSys* approach has been developed at the *Institute of Databases and Information Systems at Ulm University* that enables the creation of complex mobile data collection applications without any programming-knowledge.

The main focus was on independence from IT-Experts, flexibility, ease of operation and the support of different mobile operating systems. Furthermore, multilingualism and complex navigation through the process of data collection (i.e., displaying questions depending on already given answers) should be supported [5, 6]. Generally, the

QuestionSys approach represents a generic questionnaire system enabling mobile data collection [7]. It allows domain experts (i.e., end-users) to manage the whole *Mobile Data Collection Lifecyle* (Creation, Deployment, Execution, Analysis, Archiving), described in [5], by themselves.

For a better understanding, the architecture of *QuestionSys* is shown in Figure 2.1. It contains the following three main components.

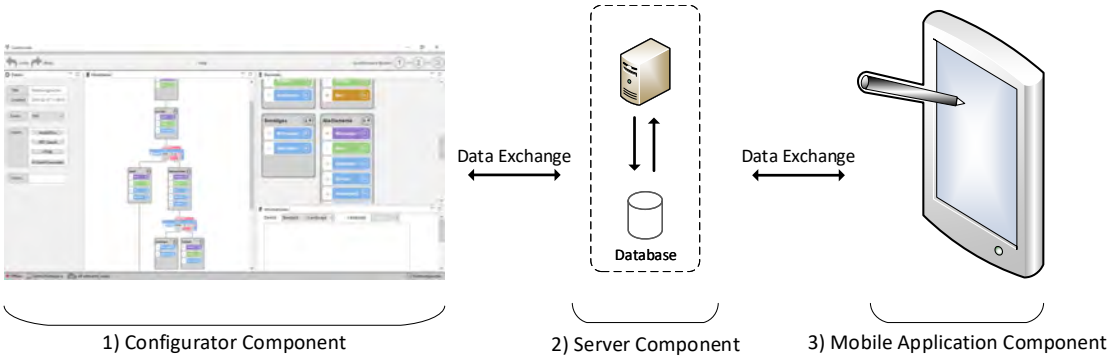


Figure 2.1: QuestionSys Framework

Configurator Component: A graphical application for defining a questionnaire (cf. Figure 2.1 ①). Typical elements of questionnaires such as *Headers, Texts or Questions* may be defined within this application. Furthermore, end-users may aggregate the latter to thematically related elements (e.g., *Pages*) as well as model a data collection instrument. Furthermore, the execution order is defined [6, 8].

Server Component: A middleware service to provide secure data exchange and to manage the collected data (cf. Figure 2.1 ②) [7].

Mobile Application Component: A mobile data collection application. It allows using smart mobile devices in order to collect data (cf. Figure 2.1 ③). Therefore, a mobile process engine is developed allowing the execution of the process model containing the logic of the instrument [6, 7].

The main focus in this thesis is on the *Configurator* component (cf. Figure 2.1 ①), which is described in more detail in the following.

2.2 Configurator Component

As mentioned before, the *Configurator* component is described in this section. This application allows end-users to create data collection instruments by defining and structuring their elements.

Figure 2.2 shows two different views of the *Configurator* component.

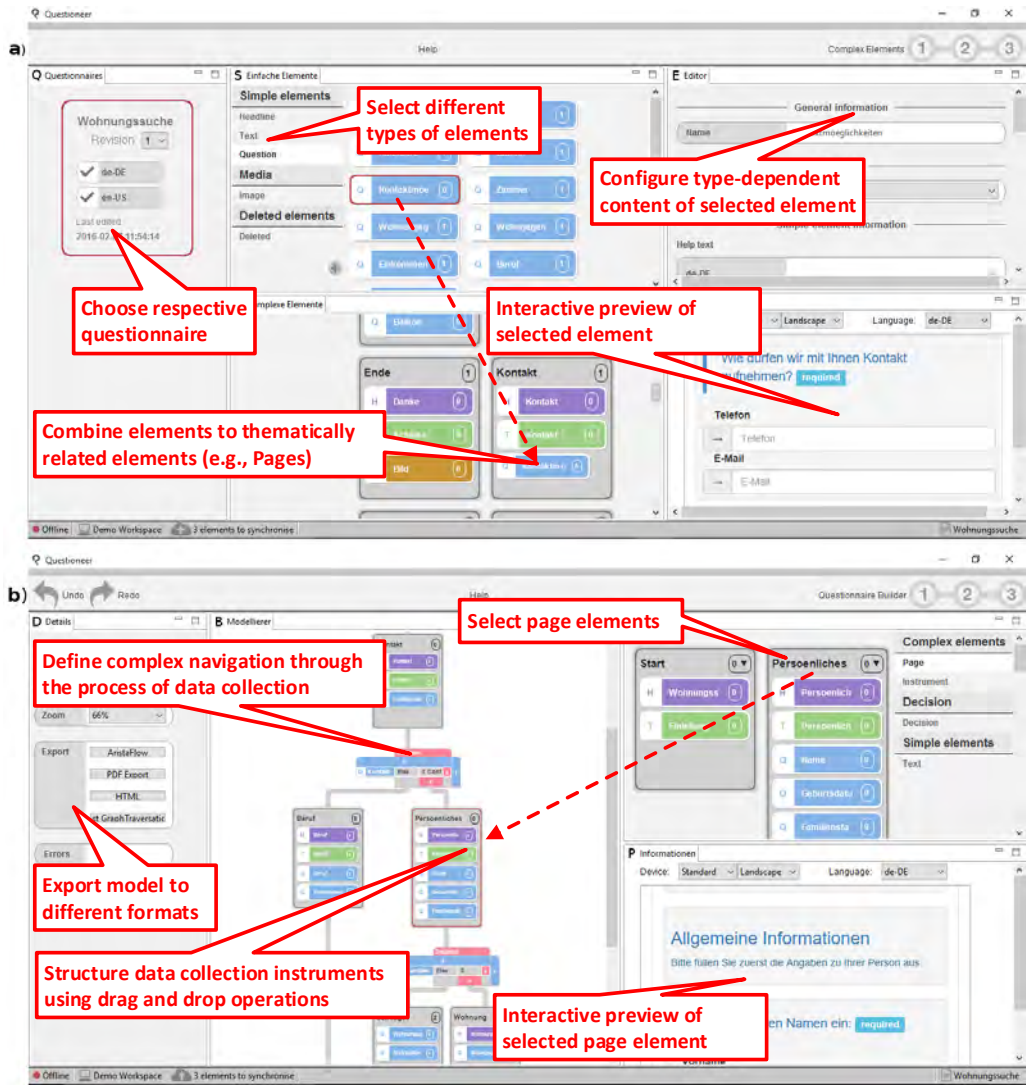


Figure 2.2: QuestionSys Configurator: a) Element View; b) Modeling View (adapted from [8])

First, the *element view* has to be described (cf. Figure 2.2 ⓐ). In the leftmost part, already created questionnaires are listed. The first step for domain experts is to choose one of the latter by double-clicking the respective one. Then, in the center-part, elements of different types (e.g., *Headlines, Texts or Questions*) may be selected. Its type-dependent content may be edited and configured in the rightmost part. This editor allows editing and managing details of the selected element, i.e., defining its name, handling multiple languages or defining a specific type of question. Additionally, versioning is supported as well, allowing to manage different instances of an element. An element has to be explicitly finalized in order to use its modifications in the following. However, finalized versions of an element can not be edited anymore. The bottom-right corner of the configurator shows an *interactive preview function*. Thus, the end-user may receive direct response when editing elements as well as getting an idea of the look & feel of the final result. Furthermore, it allows to simulate different devices and to switch between the specified languages. Another important function is shown in the bottom-center part. Elements can be combined to more complex, thematically related, elements (i.e., *Pages*) using simple drag & drop operations. The domain expert is able to add elements in arbitrary order as well as to delete elements not needed anymore. Again, the performed modifications have to be finalized [8].

Figure 2.2 ⓑ introduces the *modeling view* of the configurator. As mentioned before, domain experts may structure the defined pages (listed in the rightmost part) to a data collection instrument. Therefore, they may drag a page to the model shown in the center view and define its execution order explicitly. Furthermore, it is possible to define *decision elements* allowing to display elements depending on already given answers during the process of data collection. Finally, the configurator allows exporting the model to different formats (i.e., PDF documents, process models, HTML). The interactive preview function in the bottom-right corner, as mentioned above, is also available in this view [8].

The goal of the study presented in this thesis is to evaluate the described configurator component (*Questioneer*). In particular, the overall understanding of the (modeling) concept with respect to the complexity of the application is evaluated. The following chapter introduces the design phase of the study.

3

Experiment Definition and Planning

To evaluate the *Configurator* component, described in Section 2.2, and the overall understanding of its concept an experiment is conducted.

In this chapter the preparation of the conducted experiment is described. To fulfil the high demands on correctness of an experiment and its results recommendations are given in [4, 9]. In the following sections the arrangement of the experiment is introduced with respect to the latter (cf. Figure 3.1).

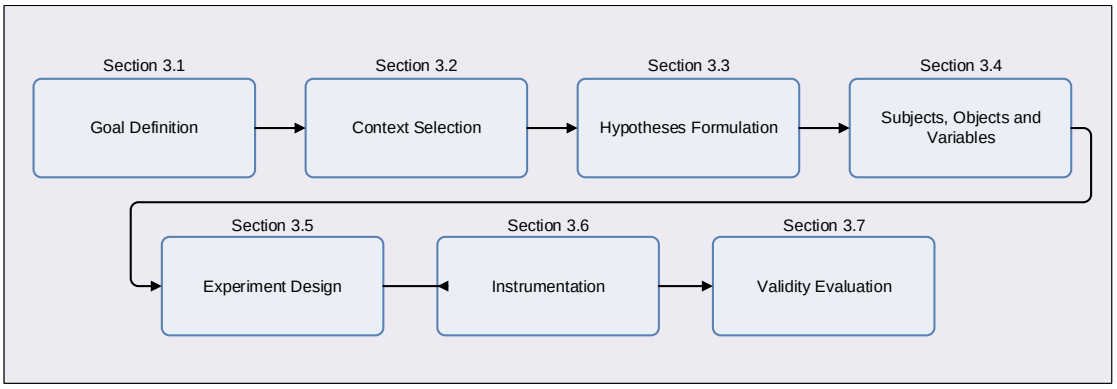


Figure 3.1: Experiment Definition and Planning

In Section 3.1 the goal of the experiment is defined. The general purpose of context selection is described in Section 3.2, while Section 3.3 introduces the established hypotheses. Section 3.4 explains the derived variables, subjects and objects. The experiment design is introduced in Section 3.5. Section 3.6 discusses the instrumentation chosen for this experiment, while Section 3.7 examines possible threats to the validity of the results.

3.1 Goal Definition

Nowadays, software engineering is mostly a complex task. Many aspects have to be considered when developing a new software application. The development process may include many people and run over a long period of time. The effort for developing a software application increases significant with the complexity of the final product.
The foundation of a new software application is a product idea. The latter is then implemented during the development process resulting in a final product [4]. There are different models for the software development process (e.g., waterfall model, V-shaped model etc.) described in general software engineering literature [10]. To stay economical competitive, enterprises have to spent plenty of money on the development process. Anyway, it is not guaranteed that end-users accept the final application. Therefore, additional quality aspects are required concerning the software application itself. According to [2, 3] the main software product quality characteristics are: functional suitability, performance efficiency, compatibility, usability, reliability, security, maintainability and portability. The focus in this thesis is on the usability of the configurator component of the QuestionSys approach. In particular, it investigates how much effort domain experts need in order to properly handle Questioneer (i.e., to solve specific tasks like creating a new questionnaire). The following fundamental research question can be established:

Do end-users (e.g., medical doctors, psychologists) understand the (modeling) concept of the software application Questioneer, with respect to the complexity of the provided application?

The best way to test the overall understanding of the concept of Questioneer is to have people using it. Therefore, a controlled software experiment is run.

Conducting an experiment needs a proper preparation to minimize or even eliminate threats affecting the outcome. Therefore, the goal definition is strongly necessary to ensure that all important aspects are considered properly. Otherwise, it may be possible that the intention of the study can not be fulfilled. The *Goal Question Metric (GQM)* described in [4, 11] helps defining the goals of an experiment in a proper way. In this particular case, it is defined as follows:

Object of Study: The *object* to be analyzed is the application software Questioneer.

Purpose: The *purpose* of this study is to evaluate the overall understanding of the (modeling) concept with respect to the complexity of the application.

Quality Focus: The main effect to be studied is the *intuitiveness of the (modeling) concept*. To measure the latter we focus on the *time* needed to solve specific tasks with different levels of difficulty and the *number of errors* made in the resulting questionnaire models.

Perspective: The *perspective* is from the developers / researchers point of view. They would like to know how much effort domain experts need in order to handle Questioneer.

Context: The experiment is conducted at the Institute of Database and Information Systems (DBIS) at Ulm University. Therefore, students and research associates are recruited as subjects.The study focuses on the configurator component of the QuestionSys approach and the resulting questionnaire models developed when solving specific tasks are observed. Each subject has to solve two tasks, hence two resulting questionnaire models.

To summarize the definition a *goal definition template* , described in [4], can be used and is shown in Table 3.1.

Analyze	the application software Questioneer
for the purpose of	software and concept evaluation
with respect to the	intuitiveness of the concept
from the point of view of	developers and researchers
in the context of	students and research associates

Table 3.1: Goal Definition Template

3.2 Context Selection

The general aim of conducting an experiment is to receive most significant results of what is investigated. Therefore, it is recommended to execute the experiment in a real

environment with professional stakeholders (in our case e.g., psychologist or medical staff). However, this procedure may involve unexpected risks (e.g., delayed delivery time of other products as a result of a time-consuming experiment). To minimize the latter it is possible to implement the experiment as an off-line experiment in a controlled environment. As subjects students may be considered. Thus, costs and time may be reduced as well as the experiment may be easier to control. Another alternative is given, when conducting the experiment as an off-line experiment in parallel to real projects. However, this option increases the costs again [4].
Within the scope of this thesis the experiment involves students and research associates. It is run as an off-line experiment in a controlled environment at the Institute of Database and Information Systems at Ulm University. For this purpose, the computer lab of the latter is prepared. The room includes twelve workstations but we decided to involve a maximum of eight subjects at the same time in order to provide enough space for each of them. Furthermore, this allows to react instantly to technical malfunctions. Since the native language of most students at Ulm University is German and, therefore, to avoid possible misunderstandings within the task descriptions, we decided to choose German as working language.

3.3 Hypotheses Formulation

Hypotheses are the foundation of research work. In general, a hypothesis describes a theoretical assumption and its core statement has to be tested. Based on the results of the hypothesis testing statistical analysis can be conducted in the following. Within the scope of the planning phase, the definition described in Section 3.1 has to be clearly stated into hypotheses. Therefore, two types of hypotheses have to be formulated [4, 12]:

A null hypothesis, H_0: Within the scope of hypothesis testing the null hypothesis describes the assumption to be tested. Therefore, this hypothesis is assumed to be true and the experimenter tries to reject it. Generally, this hypothesis indicates that there are no tendencies in the experiment setting. Possible differences in observations are only based on random reasons [4]. $H_0 : \mu_{N_{old}} = \mu_{N_{new}}$

An alternative hypothesis, H_1: The alternative hypothesis represents the opposite of the null hypothesis. Generally, it describes the experimenters prediction or assumption and can be accepted if the null hypothesis is rejected [4]. $H_1 : \mu_{N_{old}} \neq \mu_{N_{new}}$

As mentioned before (cf. Section 3.1) the experiment evaluates the overall understanding of the (modeling) concept with respect to the complexity of the software application Questioneer. In particular, it investigates how much effort domain experts need in order to handle the application properly. Therefore, we focus on prior experience in process modeling distinguished between *novices* (users with less or no experience, (μ_1)) and *experts* (users with experience, (μ_2)). Additionally, two levels of difficulty (easy (μ_{n_1}) and advanced (μ_{n_2})) of the tasks the subjects have to deal with, are considered. As a result, six hypotheses have been derived as shown in the following:

1. *Does **experience** lead to a decrease of **time** needed to solve specific tasks?*
   $H_{0,1}$: There are no significant differences regarding the time needed when solving the tasks considering the users's experience.
   $H_{0,1}$: time(μ_1) = time(μ_2)
   $H_{1,1}$: Experts are faster with respect to solving the required tasks than novices.
   $H_{1,1}$: time(μ_2) < time(μ_1)

2. *Does an increase of **difficulty** of the tasks lead to an increase of **time** needed to solve those tasks for **novices**?*
   $H_{0,2}$: For novices there are no significant differences in time needed when solving tasks with higher difficulty.
   $H_{0,2}$: time(μ_{1_1}) = time(μ_{1_2})
   $H_{1,2}$: Novices are significantly slower in solving tasks with higher difficulty.
   $H_{1,2}$: time(μ_{1_1}) < time(μ_{1_2})

3. *Does an increase of **difficulty** of the tasks lead to an increase of **time** needed to solve those tasks for **experts**?*
   $H_{0,3}$: For experts there are no significant differences in time needed when solving tasks with higher difficulty.
   $H_{0,3}$: time(μ_{2_1}) = time(μ_{2_2})

$H_{1,3}$: Experts are significantly slower in solving tasks with higher difficulty.
$H_{1,3}$: time(μ_{2_1}) < time(μ_{2_2})

4. *Does **experience** lead to a decrease of **errors** made in the resulting questionnaire model?*
   $H_{0,4}$: There are no significant differences in the number of errors made regarding the users's experience.
   $H_{0,4}$: errors(μ_1) = errors(μ_2)
   $H_{1,4}$: Experts make less errors than novices.
   $H_{1,4}$: errors(μ_2) < errors(μ_1)

5. *Does an increase of **difficulty** of the tasks lead to an increase of **errors** made in the resulting questionnaire model for **novices**?*
   $H_{0,5}$: For novices there are no significant differences in the number of errors made when solving tasks with higher difficulty.
   $H_{0,5}$: errors(μ_{1_1}) = errors(μ_{1_2})
   $H_{1,5}$: Novices make significant more errors when solving tasks with higher difficulty.
   $H_{1,5}$: errors(μ_{1_1}) < errors(μ_{1_2})

6. *Does an increase of **difficulty** of the tasks lead to an increase of **errors** made in the resulting questionnaire model for **experts**?*
   $H_{0,6}$: For experts there are no significant differences in the number of errors made when solving tasks with higher difficulty.
   $H_{0,6}$: errors(μ_{2_1}) = errors(μ_{2_2})
   $H_{1,6}$: Experts make significant more errors when solving tasks with higher difficulty.
   $H_{1,6}$: errors(μ_{2_1}) < errors(μ_{2_2})

Furthermore, hypothesis testing contains different types of risks that need to be considered. In the following the different types of risks are briefly introduced, as described in [4].

Type-I-Error: A *type-I-error* occurs if the null hypothesis is rejected even it is true. The probability can be described as follows:
$P(type - I - error) = P(H_0\ is\ rejected \mid H_0\ is\ true)$

Type-II-Error: A *type-I-error* occurs if the null hypothesis is not rejected even it is false. The probability can be described as follows:
$P(type - II - error) = P(H_0\ is\ not\ rejected \mid H_0\ is\ false)$

Power: Generally, the *power* describes the significance of a statistical test. In particular, it is the probability that a null hypothesis is rejected correctly when the alternative hypothesis is true. Therefore, choosing a statistical test with a high power as possible is recommended. The probability can be described as follows:
$Power = P(H_0\ is\ rejected \mid H_0\ is\ false) = 1 - P(type - II - error)$

3.4 Subjects, Objects and Variables

Once, the goal of the study is defined and its hypotheses are stated further details of the experimental setup have to be described. This section introduces different factors of the experiment as there are subjects participating, objects investigated and variables (independent, dependent) selected.

Selecting subjects: Here, subject means a test-person participating in an experiment. Choosing the right subjects is necessary when conducting an experiment. Achieving the most significant results would mean the whole desired population has to be investigated. Usually, it is not possible to investigate the whole desired population. Therefore, a representative selection has to be chosen in order to be able to generalize the results from the experiment. This operation also called *sampling*. Generally, there are two approaches for sampling a population, probability sampling (e.g., simple random sampling, systematic sampling, stratified random sampling) and non-probability sampling (e.g., convenience sampling, quota sampling) described in more detail in [4]. In our particular case (cf. Section 3.1), the experiment investigates the overall understanding of the (modeling) concept of a software application newly developed. Although, its main target groups should be domains where data collection in large scale is needed (e.g., medical domains), other domains may benefit from the developed configurator application as well. As a consequence, domain expertise is not necessary and, therefore, students and research associates from different courses of study as well as with different

background knowledge (from none to advanced) in process modeling were chosen using the convenience sampling method.

Selecting objects: As a next step, the objects to be investigated have to be determined. In our experiment the subjects have to solve two specific tasks with the software application Questioneer. In particular, each subject has to model a questionnaire per task. The resulting questionnaire models (exported as pictures) are the objects to be observed.

Choice of independent variables: Controllable variables in an experiment are called *independent variables*. Since they have direct effect on the dependent variable it is important considering them thoroughly. Often, another definition is found in the literature, e.g., *factors*. Changing one or more factors leads to a measurable effect in the dependant variable. If there are more than one factor in an experiment and a desired effect of changing a specific one has to be investigated, it is important to control the other factors at a fixed level. Otherwise, it is not guaranteed that the observed effect is a result of manipulating the specific one. A certain value of a factor is called *treatment* [4].
In our particular case, two factors have to be considered. On the one hand we consider the *experience* in process modeling of the subjects. Treatments are *novice* and *expert*. Therefore, we divided the subjects into two groups. Subjects with less or no prior experience in process modeling form the group of novices while subjects with advanced experience in process modeling form the group of experts. As a criterion to distinguish between novices and experts we decided to focus on the amount of process models a subject has read/analyzed respectively created/edited within the last 12 months. In our particular case, if a subject has read/analyzed more (or equal) 20 process models or has created/edited more (or equal) 10 process models within the last 12 months, advanced experience in process modeling is assumed. Thus, the participant is treated as an expert. Otherwise the subject is treated as a novice. On the other hand we consider the *level of difficulty* of the tasks the subjects have to deal with. Treatments are *easy* and *advanced* based on the complexity of the different tasks developed. As a criterion for complexity we decided to focus on the number of decision elements (cf. Section 2.2) the questionnaire models should include. No decision element means that the level of difficulty of the task is easy while one decision element means that the task has an advanced level of difficulty.

Choice of dependent variables: The next step is choosing the dependent variables. In the latter an occurring effect of a treatment, as described above, can be measured. Therefore, it is necessary to choose the right variables, often called response variables [4]. For a better understanding Figure 3.2 illustrates independent and dependent variables in an experiment.

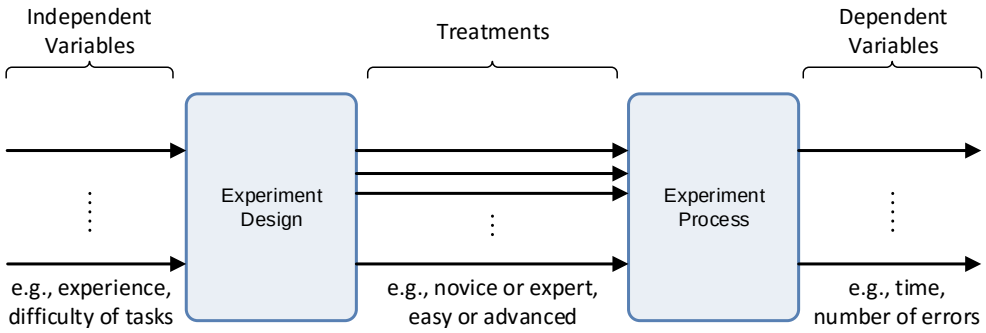


Figure 3.2: Illustration of an experiment, referred to [4]

In our particular case, we focused on two dependent variables. On the one hand, the *time* needed when solving the different tasks, as mentioned above, is considered. We assume that experience in process modeling impacts the time users need to model questionnaires. In particular, we expect that users having advanced experience in process modeling are significantly faster with respect to modeling required questionnaires than users with less or no prior experience in process modeling. Therefore, logging features (cf. Secction 3.6) regarding the time and some additional informations were implemented within Questioneer allowing to compute the overall time needed. On the other hand, the *number of errors* made in the resulting questionnaire models are evaluated. Again, we assume that experience in process modeling impacts the number of errors made by users during modeling. In particular, we expect that less errors are made by users with advanced experience in process modeling. Therefore, the resulting questionnaires are investigated and different types of errors are classified. Then, the latter are weighted in order to obtain an appropriate result (cf. Appendix A).

Summarizing the aspects, mentioned above, two independent variables (prior *experience* in process modeling and *difficulty* of the tasks) are determined.

1. **Experience** is measured by a classification into two groups:

   a) Less or no prior experience in process modeling (novices). In particular, subjects, which have read/analyzed less than 20 process models and have created/edited less than 10 process models within the last 12 months.

   b) Advanced experience in process modeling (experts). In particular, subjects, which have read/analyzed more or equal 20 process models or have created/edited more or equal 10 process models within the last 12 months.

   Hence, an ordinal scale is used to measure the experience in process modeling.

2. **Difficulty** of the tasks is measured by classifying the tasks into two classes:

   a) Easy: a questionnaire model including no decision element (cf. Section 2.2).

   b) Advanced: a questionnaire model including one decision element.

   Hence, an ordinal scale is used to measure the difficulty of the tasks to be processed.

Furthermore, two dependent variables (*time* and *errors*) are determined.

1. **Time** to complete the tasks is measured in seconds.
   Hence, a ratio scale is used.

2. **Errors** are measured by a classification of the different types of errors observed into three main groups (cf. Appendix A) regarding the different elements which are introduced in Section 2.2:

   a) Errors concerning a simple element (e.g., headers, texts or questions).

   b) Errors concerning a complex element (e.g., pages).

   c) Errors concerning a decision element.

   Hence, an ratio scale is used to measure the errors made in the resulting questionnaire models.

3.5 Experiment Design

Once the goal is set, the hypotheses are defined, the context and the variables are chosen the experiment has to be designed. The organisation and execution of an experiment is defined within the experiment design. Therefore, it is necessary to determine the experiment design as accurately as possible in order to guarantee valid and meaningful results. Three general design principles are described in [4, 13] allowing to specify the experiment design.

1. **Randomization:** The first design principle allowing to guarantee meaningful results is called *randomization*. All experimental units (e.g., objects, subjects and its assignment to the treatments) should be assigned randomly. Therefore, it allows neglecting an impact of a factor that may exists otherwise. However, in our experiment the objective of the study is to evaluate a newly developed software application. Furthermore, no comparison to another software application has taken place. Therefore, the object could not be assigned randomly to the subjects as everyone uses Questioneer to process the tasks. Convenience sampling was used in order to select the subjects (cf. Section 3.4). The allocation of the latter regarding the two groups (novices and experts) are based on the criterion described in Section 3.4 and, therefore, is given. Furthermore, the assignments to each treatment (the two levels of difficulty of the tasks) are not made randomly since everyone should process the same tasks in order to be able to compare the results.

2. **Blocking:** Another design principle is called *blocking*. Using this principle, an undesired effect that may possibly affect the subjects can be eliminated. Therefore, the effect has to be known and controllable. In our experiment, we grouped the subjects into two groups (blocks), i.e., novices and experts, both with different experience in process modeling.

3. **Balancing:** In our experiment, inter alia, we want to investigate a difference between two groups of subjects. Statistical analysis of the data can be strengthened when using an balanced design. Balanced means an equal number of subjects in each group. Also it is desirable, it is not necessary [4]. Since it was not possible to

know about the subject's prior experience in process modeling before starting the experiment, we have no balanced design in our experiment.

Figure 3.3 summarizes our experiment design. Two groups were formed (i.e., novices and experts) and each group deals with the different treatments of the factor difficulty of the tasks.

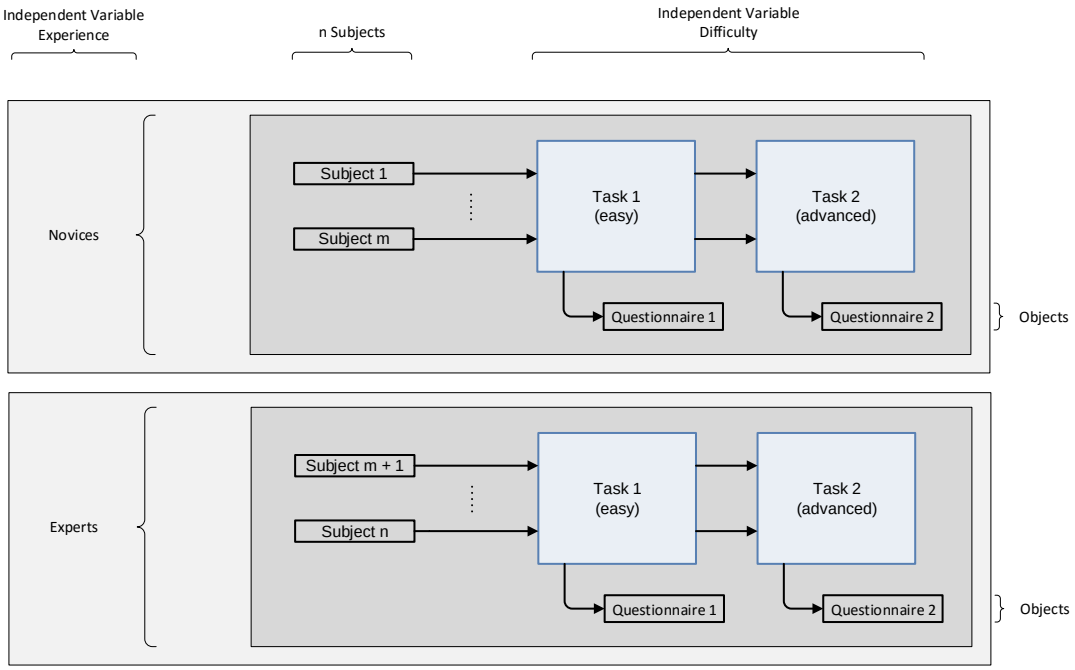


Figure 3.3: Experiment Design

3.6 Instrumentation

This section introduces the *instrumentation* needed to execute an experiment in an adequate way. Proper instrumentation is necessary to guarantee valid data. However, the instrumentation shall not affect the subjects, the data collected during the experiment nor the experiment process itself in an undesired way. Therefore, it is recommended to choose appropriate instruments for measuring data as well as guidelines for subjects during the planning and design phase [4]. In our particular case, different instruments

are determined to evaluate the software application Questioneer.
First thing to be mentioned are the implemented *logging features*. Thus, Questioneer was modified and additional features were developed to guarantee an appropriate measurement of the needed time for modeling a specific task. Additionally, each operation made by subjects when modeling the questionnaire was captured by saving a picture of the current state of the model. This allows to reproduce the single steps a subject has performed. An identification number was given to each subject allowing to associate the log files to the paper-based questionnaires which will be introduced in the following. An extract of a log file is given in Table 3.2.

Timestamp	Name	Version	Action	Prev. Page	Next Page	Custom Data
1468931965823	Willkommen	0	PAGE_INSERTED	STARTNODE	ENDNODE	
1468931974798	Ausstattung	0	PAGE_INSERTED	Willkommen	ENDNODE	
1468932142937	Decision	-1	SPLIT_INSERTED	Ausstattung	ENDNODE	
1468932152239	Decision	-1	PATH_ADDED	Ausstattung	JOINNODE	Path Index:1
1468932155351	Decision	-1	PATH_ADDED	Ausstattung	JOINNODE	Path Index:2
1468932156988	Decision	-1	PATH_DELETED	Ausstattung	JOINNODE	Path Index:2

Table 3.2: Log File Extract

Each single row of a log file represents an operation made by the subjects when modeling the given task. Therefore, for each operation different information is logged within the columns which allows reproducing the performed action. In particular, an operation consists of a *timestamp* allowing to compute the needed time. Furthermore, the *name* and the *version* number of the element are logged allowing to identify the element which is involved in the respective *action* performed. The latter is logged in the next column. This information is needed to make statements about what happens with the element (i.e., is it inserted or deleted). In the next two columns (Prev. Page, Next Page) the range within the action is performed is logged. *Previous page* specifies the name of the previous element while *next page* specifies the name of the following element. Finally, *custom data* (e.g., the path index number) is logged allowing to reproduce modifications on decision elements.
The second measurement instrument is a sheet to evaluate the *errors* made by the subjects during modeling the questionnaires within Questioneer. Appendix A shows the different types of errors made as well as how they are classified as mentioned in Section 3.4. The final models were investigated by expecting the exported pictures of the models (and the captured pictures of each operation) and documenting the errors within the

error evaluation sheet.

At the beginning of the experiment the subjects received an *introduction* to Questioneer. The latter was provided as short *tutorial* showing the main components of Questioneer (cf. Section 2.2). Different examples how to handle different tasks were introduced to the subjects. This tutorial took about five to ten minutes and the subjects were able to ask questions if needed.

As an additional help for dealing with the tasks, different *screencasts* developed were provided. The latter explain the single components of Questioneer in more detail and were placed within a study folder on each workstation. The subjects were allowed to take a look at the screencasts during modeling.

After the introduction, a *questionnaire* has to be answered regarding some demographic questions (cf. Table 3.3). The latter allow to specify the experience in process modeling of the subjects as well as to collect some background informations, i.e., their course of study, their experience in using Questioneer or their graduation level.

Question	**Possible Answers**
Welche berufliche Ausbildung trifft am ehesten auf Sie zu?	beruf. Ausbildungsstand[1]
Wie hoch ist Ihre aktuelle Anzahl an Ausbildungsjahren (inkl. Grundschule):	benutzerdefinierter Text
Welchen höchsten Bildungsabschluss haben Sie?	Bildungsabschluss[2]
Geben Sie Ihr Geschlecht an:	Männlich, Weiblich, Anderes
Studiengang:	benutzerdefinierter Text
Insgesamt bin ich mit "Questioneer" sehr vertraut.	7-Punkte Likert Skala[3]
Ich fühle mich im Verstehen von Fragebogen-Modellen, die mit "Questioneer" erstellt wurden, sehr sicher.	7-Punkte Likert Skala[3]
Ich fühle mich im Benutzen von Questioneer für Fragebogen-Modellierung sehr kompetent	7-Punkte Likert Skala[3]
Vor wie vielen Jahren haben Sie mit der Modellierung von Prozessmodellen begonnen?	benutzerdefinierter Text
Wie viele Prozessmodelle haben Sie innerhalb der letzten 12 Monate analysiert oder gelesen?	benutzerdefinierter Text
Wie viele Prozessmodelle haben Sie innerhalb der letzten 12 Monate erstellt oder bearbeitet?	benutzerdefinierter Text
Wie viele Aktivitäten haben alle diese Modelle im Durchschnitt?	benutzerdefinierter Text
Wie viele Tage formale Ausbildung zu Prozessmodellierung haben Sie in den letzten 12 Monaten erhalten?	benutzerdefinierter Textt
Wie viele Tage haben Sie in den letzten 12 Monaten mit dem Selbststudium zu Prozessmodellierung verbracht?	benutzerdefinierter Text
Vor wie vielen Monaten haben Sie begonnen "Questioneer" zu benutzen?	benutzerdefinierter Text

Table 3.3: Demographic Questionnaire, adapted from [14]

Furthermore, the tasks to be solved with Questioneer have to be developed. Therefore, two different tasks have been developed, each with a different level of difficulty as mentioned in Section 3.4. Appendix B shows the different tasks the subjects have to deal with.

[1] Auszubildende[r]/Student[in], abgeschlossene Berufsausbildung (Lehre), abgeschlossene Ausbildung an einer Meister- oder Technikerschule, Akademiker[in], Sonstiges und zwar:

[2] ohne Abschluss, Hauptschulabschluss oder Volksschulabschluss, Realschulabschluss (Mittlere Reife), Fachhochschulreife, Hochschulreife (Abitur), Fachhochschulabschluss, Hochschulabschluss, Sonstiger Abschluss und zwar:

[3] trifft völlig zu, trifft zu, trifft eher zu, neutral, trifft eher nicht zu, trifft nicht zu, trifft gar nicht zu

To ease the process of modeling a questionnaire, a workspace within Questioneer has been prepared. The workspace, in turn, includes two different predefined questionnaires, one for each task. Within these questionnaires simple and complex elements (cf. Section 2.2) are predefined allowing the subjects to start modeling right away and use the given elements when solving the tasks.
Then, after each task the subjects had to answer a brief questionnaire regarding the mental effort they needed to solve the task as well as if they had to take a look at the screencasts. Table 3.4 shows the mental effort questionnaire with respect to a single task.

Question	Possible Answer
Der mentale Aufwand zur Erstellung des gesamten Modells war sehr hoch.	7-Punkte Likert Skala[3]
Der mentale Aufwand zur Umsetzung der expliziten Änderung war sehr hoch.	7-Punkte Likert Skala[3]
Denken Sie, Sie konnten die Aufgabe korrekt lösen?	7-Punkte Likert Skala[3]
Mussten Sie einen oder mehrere Screencasts anschauen um die Aufgabe zu lösen?	Ja / Nein

Table 3.4: Mental Effort Questionnaire per Task

Considering the evaluation of Questioneer, an additional comprehension questionnaire (cf. Appendix C) was developed to investigate if the subjects have understood the most important aspects of Questioneer. For this purpose twelve questions were determined. Finally, another questionnaire (cf. Table 3.5) has been developed regarding the mental effort again. Thus, additional information can be collected in order to specify how the subjects perceive the quality of their developed questionnaire models.

Question	Possible Answer
Stimmen Ihre Modelle aus Ihrer Sicht mit den in den Aufgaben beschriebenen Fragebögen überein?	7-Punkte Likert Skala[3]
Gibt es signifikante Aspekte, die in Ihren Modellen fehlen?	7-Punkte Likert Skala[3]
Beschreiben Ihre Modelle exakt die beschriebene Logik und Semantik der vorgegebenen Fragebögen?	7-Punkte Likert Skala[3]
Gibt es, Ihrer Meinung nach, gravierende Fehler in Ihren Modellen?	7-Punkte Likert Skala[3]
Hätten sie nochmal die Gelegenheit dazu, würden Sie an Ihren Modellen etwas modifizieren?	7-Punkte Likert Skala[3]

Table 3.5: Mental Effort Questionnaire Final

3.7 Validity Evaluation

This section discusses the validity of the experiment results. There are certain factors to be considered that may threaten the outcome of an experiment. Therefore, considering

the question of validity already during the planning phase reduces the probability of invalid results. [4, 12, 15] introduce four types of validity that are important to be taken into account when planning an experiment. Figure 3.4 summarizes the four types of validity described in the following. Additionally, examples of threats to the respective validity are introduced according to [4, 12].

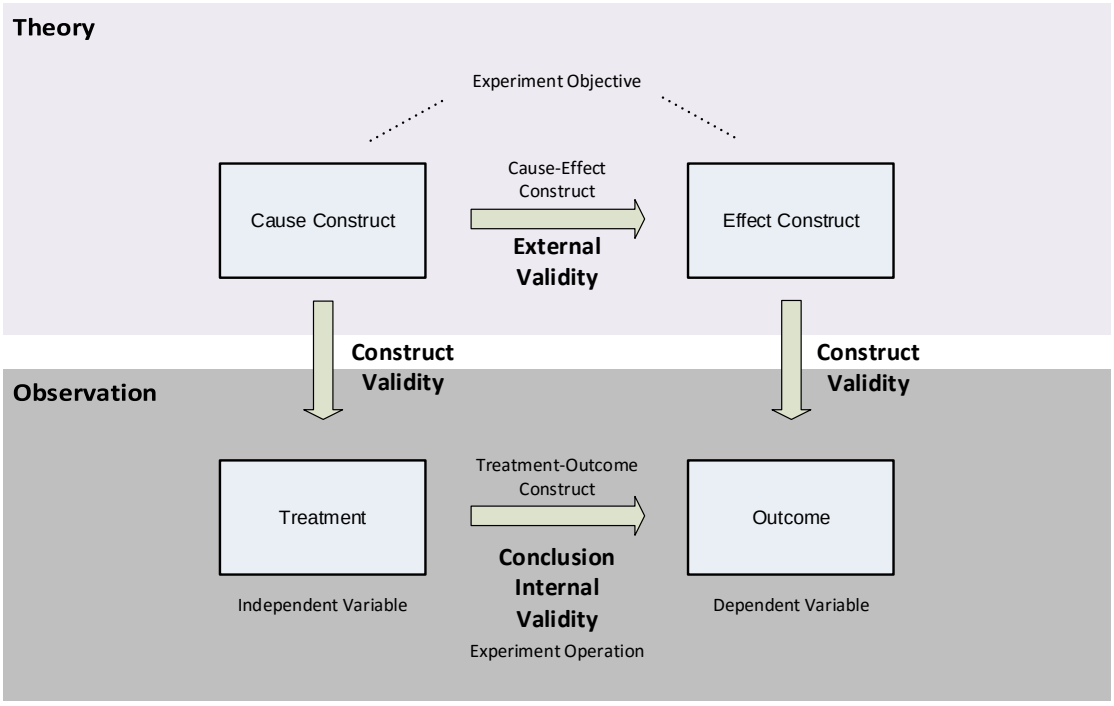


Figure 3.4: Validity Types, adapted from [4, 12]

Conclusion Validity addresses the question: Is there a relationship between the treatment and the outcome? In our case, for example, it is concerned with the relationship between the user's experience in process modeling and the time needed to solve a specific task. Two examples of threats to the conclusion validity are:
Low reliability of measures: The first example of violating the validity of the experiment results depends on the reliability of the measures. In other words, factors like bad instrumentation, bad formulation of questions or a bad design can affect the measures in an undesired way. A proper instrumentation as mentioned in Section 3.6 is important in order to enable a repetition of the study leading to same results.

Low statistical power: Another important aspect to be considered in the context of conclusion validity is to use a statistical test with a high power as possible. In Section 3.3 power is described as the probability that a null hypothesis is rejected correctly when the alternative hypothesis is true.

Internal Validity deals with the question: If there is a relationship between the treatment and the outcome, are the effects caused by the treatment itself? For instance, is an increasing time a result caused by the user's experience in process modeling or a result caused by another undesired factor? Two examples of threats to the internal validity are:
History: This threat means that there may be an imbalance between the different experiment sessions. In other words, when conducting the experiment at different times historical events (i.e., an examination or a festival the day before) can influence the outcome of the experiment. Our sessions have taken place on weekday afternoon during semester breaks to minimize this threat.
Maturation: To minimize the risk of invalid results maturation has to be taken into account. The subjects may lose motivation over the period of time the experiment process takes place. They may get bored or tired, i.e., when solving the tasks. Therefore, an adequate duration of the experiment should be chosen as well as varied tasks. The duration of our experiment process lasts about 60 minutes and can be stated as an adequate time. Half of the time was needed to fill out the questionnaires (e.g., demographic, mental effort and comprehension questionnaires) and to receive the introduction, the other half was needed to solve the given tasks.

Construct Validity handles the question: If there is a causal relationship, does the treatment reflects the cause construct well and does the outcome reflects the effect construct well? For instance, does the user's experience in process modeling reflect our theoretical definition and does the time measured reflect the time we wanted to measure? Two examples of threats to the construct validity are:
Interaction of different treatments: If the subjects take part in other studies as well, it might be that the observed results are a consequence of a treatment regarding to a different but similar study as the required one. Since Questioneer is a newly developed software application, the probability of taken part in a similar study is reduced to a minimum.

Hypothesis guessing: When participants trying to guess the purpose of the study it might be possible that they react differently. Additionally, they may figure out how the purpose should be measured. Therefore, the experiment results may be falsified. In our study, the subjects only know the intention, i.e., to test a newly developed software application, but not the purpose, namely the evaluation of the intuitiveness of the (modeling) concept, of the study.

External Validity addresses the question about generalization. In particular, if there is a causal relationship between the cause construct and the effect construct, may the results be generalized? In other words, can the results be generalized to another context (i.e., other people, other places and other times)? There are three major threats concerning the external validity:
No representative selection of subjects: Considering the generalization the selection of a representative group of people is important to draw meaningful conclusions about the experimental results. As mentioned above, domains where data collection in large scale is needed (e.g., the medical domain) should be the main target groups. However, other domains may benefit from the framework as well. Consequently, domain expertise is irrelevant and students as well as research associates can be taken instead. Therefore, the results should be transferable to other populations as well.
No representative experimental setting: Another threat to be mentioned is the choice of an adequate environment that is representative. Our study has taken place in a computer lab but could have been conducted in every other environment instead. The only requirement is that Questioneer is runnable on the chosen workstation.
No representative choice of time of the study: Finally, the time of the study has to be considered. To generalize the results of the experiment, the latter have to be independent from the particular time the experiment is conducted.

Further threats concerning the mentioned validities are described in [4, 12].

This chapter has introduced the planning and definition phase of our experiment. Therefore, the goal has been set, the context has been described, the hypotheses have been stated and the variables needed have been introduced. Furthermore, the organisation and execution of the experiment have been defined within the experiment design. Addi-

tionally, the instrumentation needed for executing the experiment has been described. Finally, different threats to the validity of the experiment have been discussed. As a next step, the experiment has to be executed. The following chapter introduces the operational process of executing the experiment. Therefore, the preparation, the execution itself and the validation of the data obtained by the experiment are described.

4

Experiment Operation

This chapter describes the final experiment operation in more detail. After defining, planning and designing the experiment it has to be executed and relevant data has to be collected. For this purpose, this chapter introduces three steps of the experiment operation: *preparation*, *execution* and *data validation*. An overview of this chapter is given in Figure 4.1.

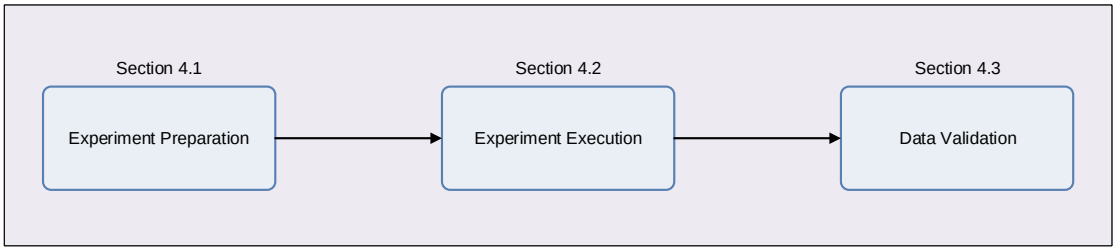


Figure 4.1: Experiment Operation

Section 4.1 introduces of what has to be done in order to guarantee an efficient and smooth experiment execution. The latter is described in Section 4.2, while Section 4.3 discusses data validation.

4.1 Experiment Preparation

A proper experiment preparation guarantees a frictionless experiment execution. For this purpose, it is necessary to select and inform possible subjects in an adequate way as well as to prepare the required material (i.e., a declaration of consent, different

questionnaires, the software application that should be investigated etc.) needed in order to execute the experiment [4].
As mentioned above, as participants of the experiment we decided to choose students and research associates. Thus, voluntaries and persons with general interests in software engineering or research work were asked to join. All subjects had to agree to participate in the experiment. For this purpose, they were informed about the research objective and obtained a declaration of consent which they had to sign. They only knew what is the intention of the experiment but not its purpose and how we investigated the latter. Within the declaration of consent, anonymity and discrete handling of sensitive data was guaranteed. The only inducement to attract participants was to offer them a candy at the end of the experiment. Furthermore, all necessary material has been designed and the software application has been prepared. Therefore, the mentioned workspace, questionnaires, task sheets and screencasts have been developed as well as the logging features have been implemented (cf. Section 3.6). Additionally, we designed a flowchart to get an overview of how the experiment should be executed. This flowchart is shown in Figure 4.2 and is described in more detail in Section 4.2.
The next step after defining and planning the experiment was its execution. Thus, pilot studies were performed to test if the desired experiment execution works as it should. As a result of two pilot studies problems during their execution, in the design as well as in the task descriptions, were figured out and improved. Furthermore, the evaluation sheet for errors (cf. Section 3.6 and Appendix A) has been developed.

4.2 Experiment Execution

This section introduces *how* the experiment was executed. First, the experiment environment has to be described. The experiment execution was proceeded at the computer lab of the Institute of Database and Information Systems at Ulm University. Therefore, a maximum of eight subjects could participate at the same time. The execution has taken place over four weeks to achieve an adequate number of participants. Thus, two sessions a day have taken place. The session duration lasts about 60 minutes. Figure 4.2 shows the running of an experiment session. Furthermore, it illustrates the allocation

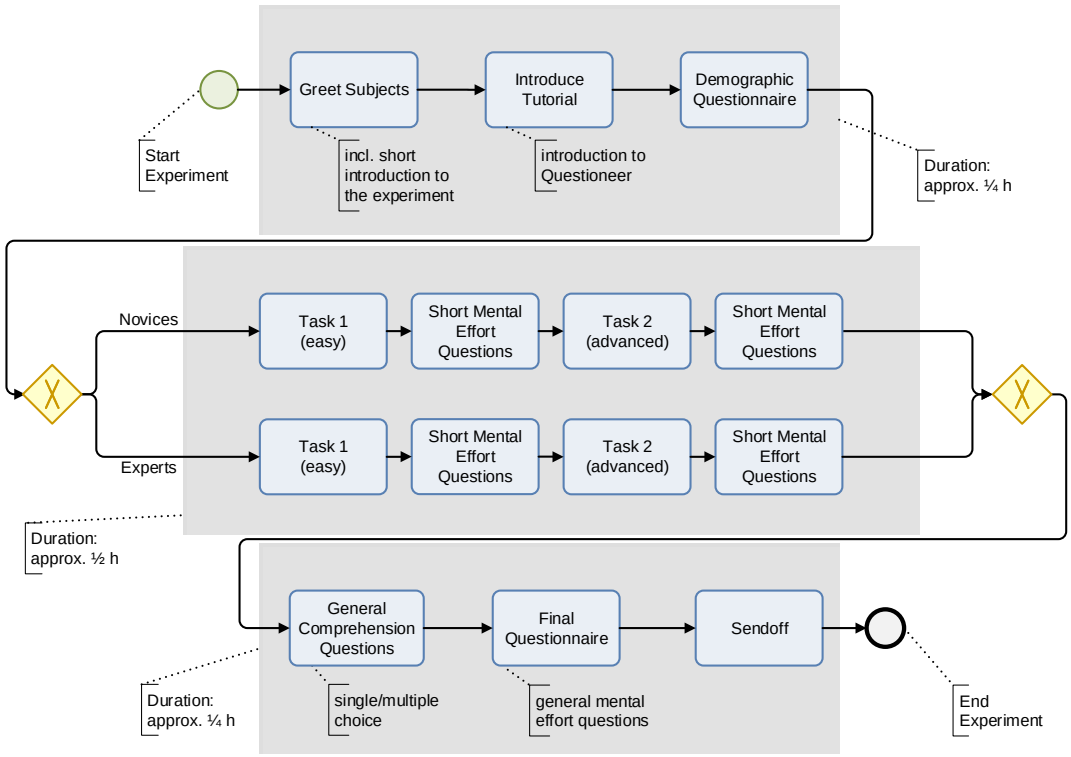


Figure 4.2: Experiment Execution

of the two investigated groups (novices and experts).

Before the beginning of each experiment procedure, the needed workstations were prepared by the staff members. For this purpose, a study folder was placed on the desktop of each workstation containing the prepared software application as well as the screencasts (cf. Section 3.6). Furthermore, all worksheets (including the declaration of consent, the different questionnaires and task sheets, cf. Section 3.6) for each subject were placed beside the respective workstation. The procedure itself (cf. Figure 4.2) can be described as follows:

The experiment procedure starts with welcoming the subjects. After that, the objective of the study is introduced and the procedure is explained briefly. During the introduction the subjects are able to ask questions if something is unclear. Furthermore, the subjects have to sign a declaration of consent. After greeting, a short introduction to the Questioneer application is given. For this purpose, a live tutorial is held by presenting

the respective software application and its most important components using a beamer. Therefore, it is guaranteed that each subject has the same background informations before using Questioneer. Following this introduction, the subjects have to fill out the demographic questionnaire (cf. Table 3.3) regarding some personal questions and their experience in process modeling. Afterwards, the subjects have to process the first task (cf. Appendix B) within Questioneer followed by the questions regarding their mental effort (cf. Table 3.4) while handling this task. As a next step, the subjects have to work on the second task (cf. Appendix B) in the same way as on the first task. Again, mental effort questions have to be answered afterwards. Subsequently, a comprehension questionnaire (cf. Appendix C) has to be filled out regarding some fundamental questions about Questioneer and its functionality. Finally, a last questionnaire has to be filled out regarding some more mental effort questions summarizing the perceived quality of the developed questionnaire models (cf. Table 3.5). At the end, we expressed our thanks to the subjects for participating in the experiment by handing them out a candy.

After each experiment procedure, the data of each workstation (created by the logging features, cf. Section 3.6) was collected in order to analyze and evaluate it in the following. Finally, the workstations were prepared for the next session, as mentioned above.

4.3 Data Validation

This section describes how the data collected during the experiment is handled with respect to correctness and validity. After conducting an experiment, the data collected must be checked regarding its correctness. Referring to [4] the experimenter has to ensure, that the participants have understood the tasks and questionnaires and, therefore, processed and filled them out properly. Furthermore, invalid data due to non-serious participation in the experiment has to be detected and removed.
After conducting our experiment, data from 44 subjects was collected. Unfortunately, data from two subjects needed to be removed due to invalidity or at least reasonable doubts regarding the correctness which may be a result of non-serious participation. The following reasons for removing the two subjects can be mentioned:

- One subject did not correctly follow the instructions given on the task-sheets and canceled the experiment while working on task two. Therefore, the results obtained from this subject are significantly different from the other ones and may affect the outcome.
- The time needed to solve the tasks of one subject differ significantly from the other ones. In particular, the time needed to solve task two was significantly higher than the other ones. This may be a result of non-serious participation. Therefore, the data was removed in order to guarantee valid data.

After removing the two subjects, data of 42 participants is left. Figure 4.3 ⓐ shows the distribution regarding their profession. We investigate 32 students (or apprentices), 9 research associates and 1 student with prior completed vocational training. Based on the prior experience in process modeling we divided the 42 subjects into two groups (cf. Figure 4.3 ⓑ). Applying our criterion (as discussed in Section 3.4) results in 24 novices and 18 experts. Figure 4.3 ⓒ shows the gender distribution. In total, we investigate 8 females and 34 males. Furthermore, the overall allocation of the courses of study for students as well as for research associates is presented in Figure 4.3 ⓓ. The distribution shows 7 different courses of study (few of them with different degree, i.e., Computer Science Dipl. or Computer Science M.Sc.) with a main focus on computer science and media informatics. Finally, Figure 4.3 ⓔ shows the distribution of the overall years of education the subjects had received at this point. We can observe that most participants had received 15 to 19 years of education. Appendix D shows the full data set of the demographic survey.

Since there are no special prerequisites to take part in the experiment (except prior experience in process modeling for experts) we may conclude that the obtained data from this 42 subjects is valid with respect to the goal of the study. As a conclusion, the data of the participating subjects can be used for further statistical analysis and interpretation which is introduced in the following chapter.

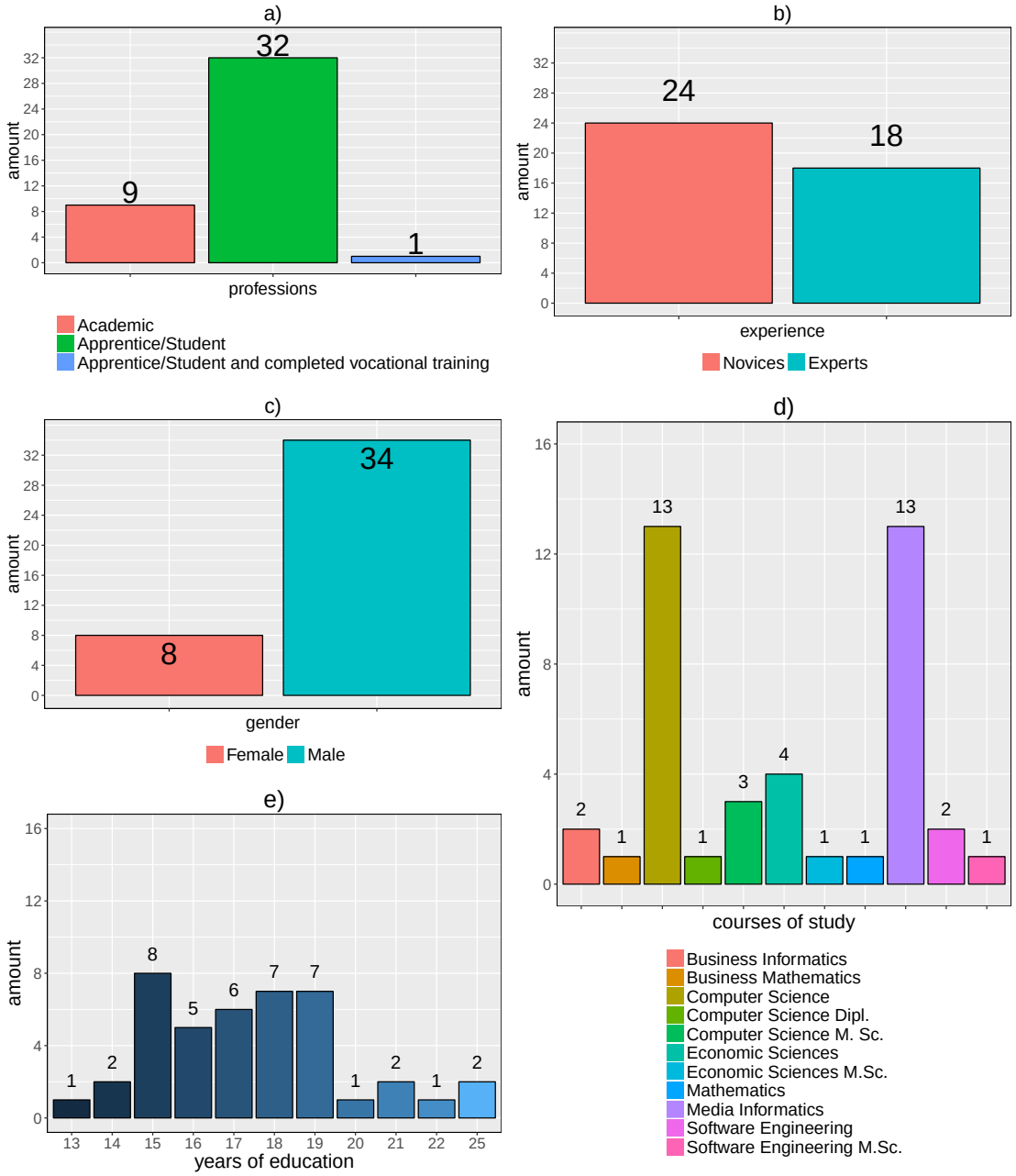


Figure 4.3: Demographic Distribution

5

Experiment Analysis and Interpretation

After conducting the experiment, the data collected has to be evaluated in order to draw meaningful conclusions. This chapter discusses how the analysis and interpretation of the data collected was achieved (cf. Figure 5.1). In Section 5.1 the data is visualized by descriptive statistics in order to get an overall understanding about its distribution. Furthermore, Section 5.2 describes data set reduction in order to get a valid data set for further statistical analysis. Hypothesis testing is introduced in Section 5.3.

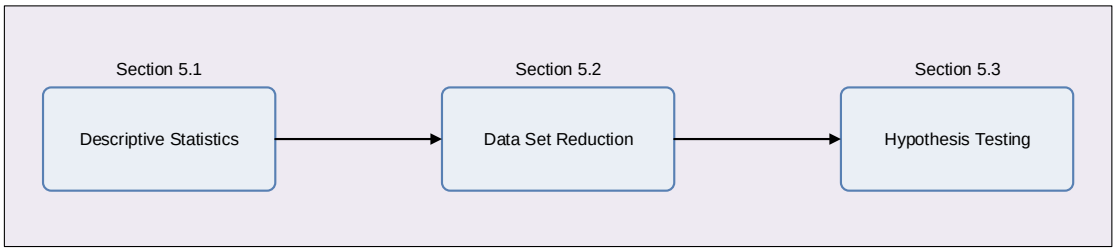


Figure 5.1: Experiment Analysis and Interpretation

5.1 Descriptive Statistics

Descriptive statistics may be useful for a better understanding of the results obtained during an experiment. The may help to describe and visualize the data in order to get an impression of its distribution. Furthermore, interesting aspects of the data can be figured out and presented to the observer in order to get a fundamental understanding for further statistical analysis. Therefore, extraordinary or at least false data points can

be identified which may negatively affect the conclusions drawn from the experimental results. Graphically presentations help to visualize possible tendencies within the data set or give an overview about its dispersion [4, 12].

This section introduces different descriptive statistics in order to get first impressions of the data obtained during the experiment (cf. Appendix E). In particular, tables consisting of median values are used to describe the obtained results. The median, as a measure of central tendency, is described as the middle value of a data set, which means it splits a list of data points into two halves of equal size. One half represents the data points that are lower than the median, the other half represents the data points that are higher than the median [4]. To visualize the results, inter alia, box plots are used. Those plots graphically represent the dispersion of a given data set. The main part of a box plot is a box representing the range of the middle 50% of the data. Its length, also called interquartile range (IQR), is computed as distance of the lower quartile (Q_1, representing the 25%-percentile) and the upper quartile (Q_3, representing the 75%-percentile): $IQR = Q_3 - Q_1$. The thicker line within the box describes the median, the lines outside the box are so-called whiskers. The latter are computed as follows: the lower whisker lw is $Q_1 - 1.5 * IQR$ and the upper whisker uw is $Q_3 + 1.5 * IQR$. The length of the whiskers can differ because their values are trimmed to the nearest data points within the calculated range instead of using the exact values computed. Box plots are good for illustrating the dispersion of a data set as well as to identify so-called outliers. The latter are data points outside the range between the lower and upper whiskers [4, 16, 17]. Furthermore, bar plots and histograms are used in order to visualize the obtained data, to gain a more precisely overview about its distribution and, therefore, to get an better understanding of the results. Especially histograms can be helpful in order to gain first informations whether a data set is normal distributed or not, since this information is necessary for further analysis. They consist of bars representing the frequencies of the different data values of a specific variable. Further tests for normal distribution (i.e., Shapiro–Wilk test [18]) can then be provided to strengthen the assumption. Additionally, matrix plots including scatter plots are introduced in order to illustrate dependencies between different calculated variables.

First thing we consider, are the total times (in seconds) the subjects needed to solve the

different tasks. Table 5.1 and Figure 5.2 show the results for novices and experts with respect to the total time in seconds each group needed to solve the given tasks.

Group	Task 1 (easy)	Task 2 (advanced)	Task 1 + 2
Novices	528.61	620.97	539.75
Experts	601.08	625.98	625.98

Table 5.1: Total Time (Median)

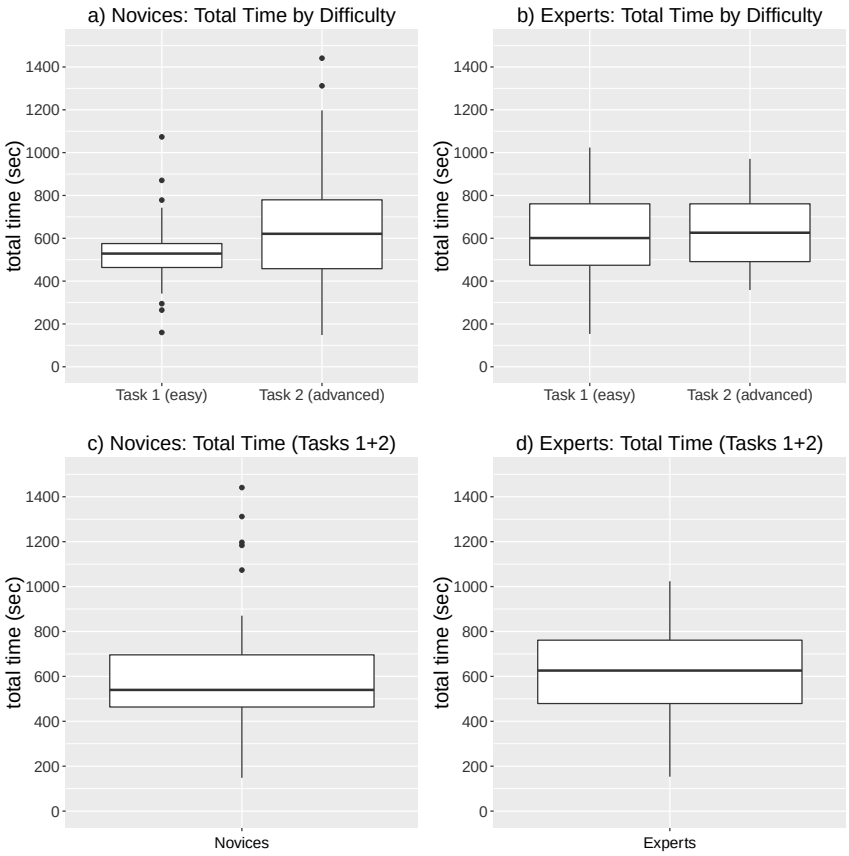


Figure 5.2: Total Time (Median)

Considering the total times (cf. Table 5.1 and Figure 5.2), novices need less time (median of 528.61 seconds) than experts (median of 601.08 seconds) when modeling task 1. However, when modeling task 2 the median values for the total time of each group barely differ (novices (620.97 seconds), experts (625.98 seconds)). Furthermore, novices as

well as experts need less time modeling task 1 than modeling task 2. This may be explained by the fact, that the complexity of task 2 is higher than the complexity of task 1. However, the difference for novices (92.36 seconds) is higher than the difference for experts (24.90 seconds) comparing the tasks, namely 3.7 times higher. Column three of Table 5.1 shows the results for each group combining the tasks. Therefore, novices need less time (median of 539.75 seconds) than experts (median of 625.98 seconds) when modeling both tasks.
Figure 5.2 presents different box plots as mentioned above. Figure 5.2 ⓐ shows the results for novices, Figure 5.2 ⓑ the results for experts separately for each task while Figure 5.2 ⓒ and ⓓ combine the results of both tasks for each group. For the group of experts, we can observe that the IQRs (size of the boxes) barley differ comparing the two tasks while there is a clear difference for those of the group of novices. Therefore, the dispersion of the middle 50% of the data of task 1 for novices is lower than the one of task 2 and even lower than the one of task 1 for experts. When comparing task 2 of each group the dispersion of the middle 50% of the data is nearly the same. Furthermore outliers can be observed within the group of novices (cf. 5.2 ⓐ and ⓒ), in particular, six outliers for the first task and two outliers for the second task.

Second thing to consider is the number of errors the subjects made when modeling the different tasks. Table 5.2 and Figure 5.3 present the results for novices and experts with respect to the number of errors made in the resulting questionnaire models.

Group	**Task 1 (easy)**	**Task 2 (advanced)**	**Task 1 + 2**
Novices	4	1	2
Experts	1	1	1

Table 5.2: Number of Errors (Median)

As shown in Table 5.2 and Figure 5.3 experts make less errors (median of 1 error) than novices (median of 4 errors) when modeling task 1. With respect to task 2 the values for the number of errors do not differ (median of 1 error). Furthermore, for experts, no difference in the number of errors can be observed considering the two different tasks (median of 1 error). However, for novices, a difference within the different tasks can be observed (task 1 (median of 4 errors), task 2 (median of 1 error)). This may be

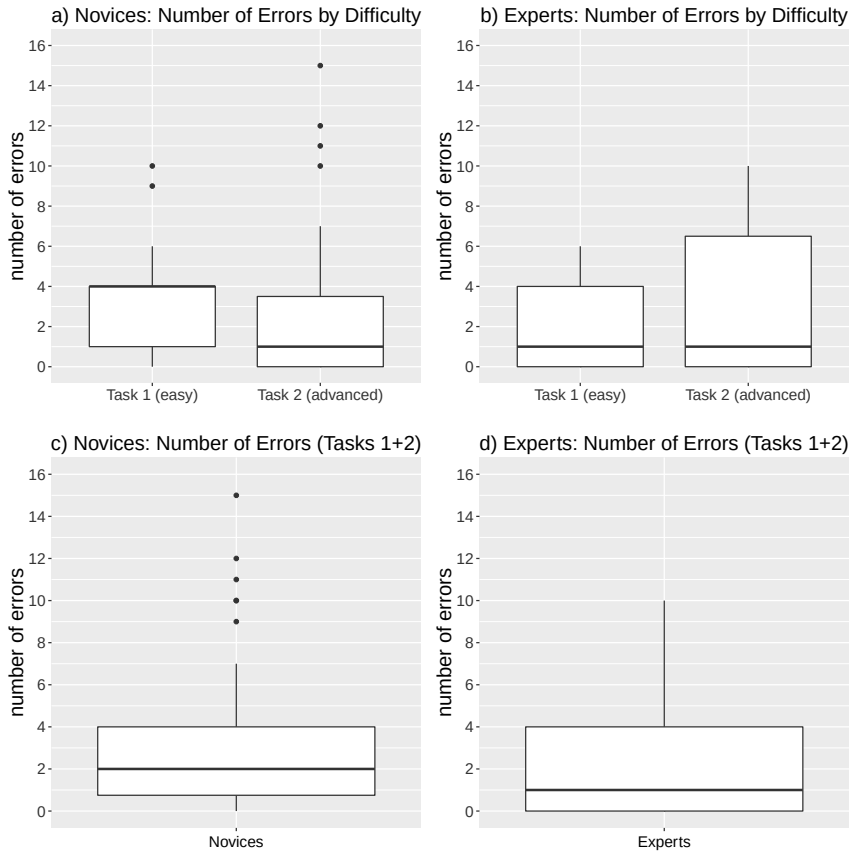


Figure 5.3: Number of Errors (Median)

explained by the fact, that novices have less or no prior experience in process modeling when processing task 1. Considering the results of the total time (cf. Table 5.1), we can observe, that novices are faster but make more errors than experts when modeling task 1. Column three of Table 5.2 shows the results for each group combining the tasks. Therefore, novices make more errors (median of 2 errors) than experts (median of 1 error) when modeling both tasks.

Box plots are presented in Figure 5.3. While Figure 5.3 ⓐ and ⓑ show the results for each task and group, Figure 5.3 ⓒ and ⓓ, again, combine the results of both tasks for the respective group. Considering the IQR for novices, we can observe an IQR of 3 for the first task and an IQR of 3.5 for the second task which means there is barley a difference regarding the dispersion of the middle 50% of the data. However, for experts

we can observe an IQR of 4 for task 1 and an IQR of 6.5 for task 2. Therefore, the difference regarding the dispersion is higher in this group. Comparing the two groups, one can notice that the dispersion of the middle 50% of the data for novices is fewer than the one for experts with respect to each task. Furthermore outliers can be observed within the group of novices (cf. Figure 5.3 ⓐ and ⓒ), in particular, two outliers for the first task and four outliers for the second task.

To gain an even better understanding of the data, bar plots are presented in Figure 5.4

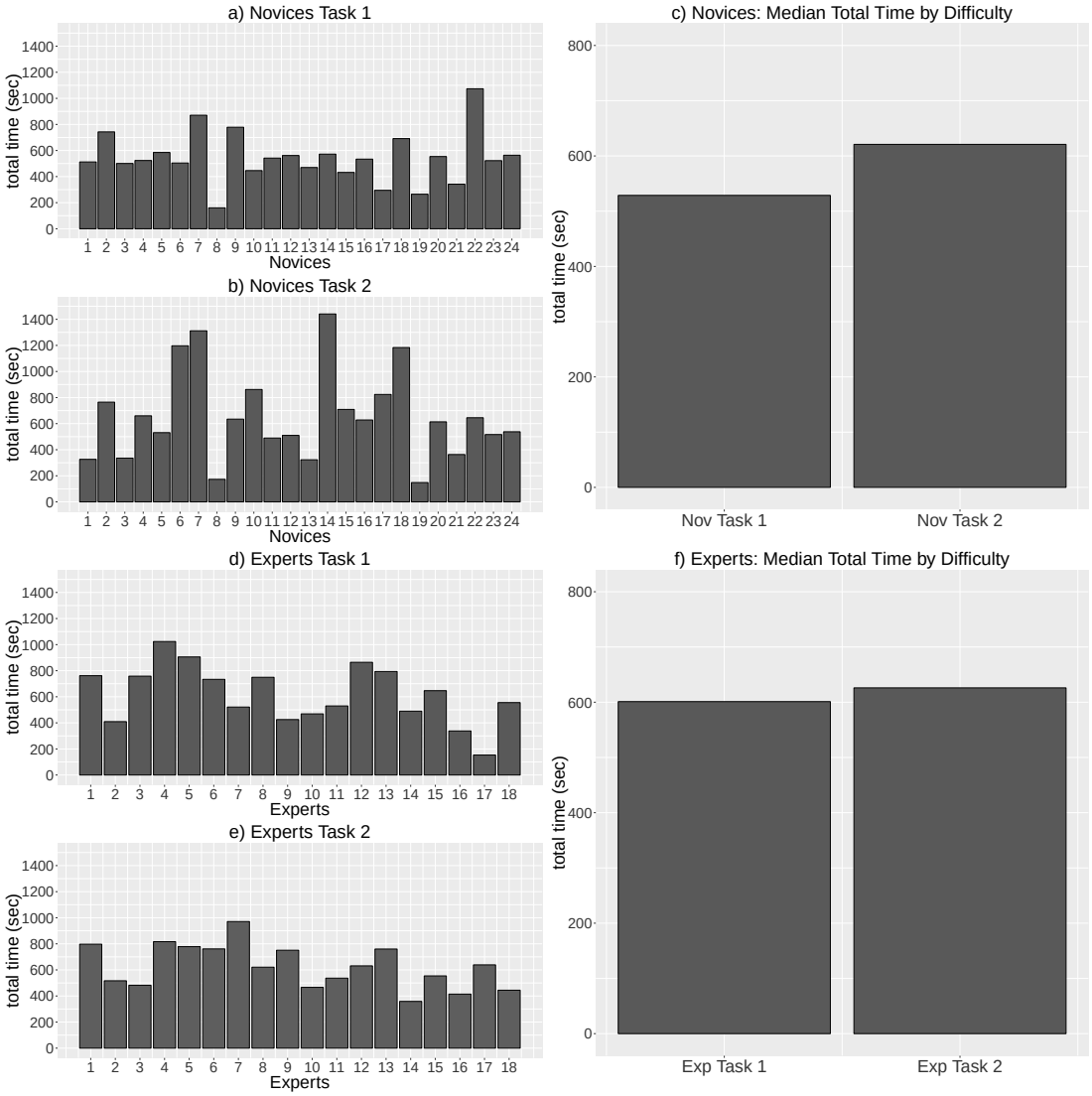


Figure 5.4: Summary Total Time

and Figure 5.5. While Figure 5.4 summarizes the data of total time the subjects needed to solve each task, Figure 5.5 shows a summary of the number of errors they made in the resulting questionnaire models.

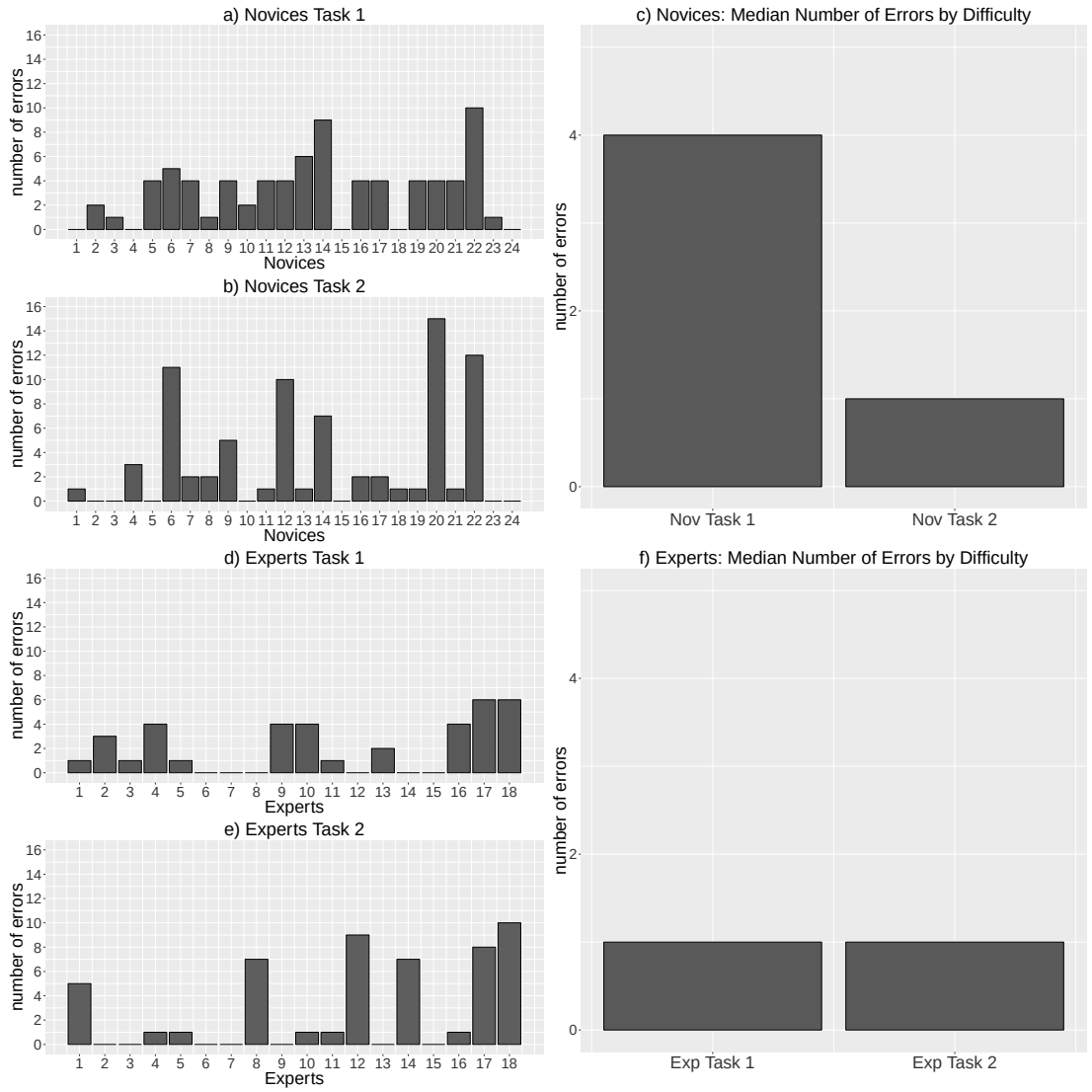


Figure 5.5: Summary Number of Errors

In particular, Figure 5.4 ⓐ and ⓑ present the total times for each single novice separated in task 1 and task 2. On the right side (cf. Figure 5.4 ⓒ), the median values for the group of novices regarding the single tasks are visualized. Figure 5.4 ⓓ and ⓔ show

the total times for each single expert, again, separated for each task while Figure 5.4 ⓕ summarizes the median values for the group of experts per task.
Figure 5.5 ⓐ and ⓑ present the number of errors for each novice separated by task while the median values for each task are shown in Figure 5.5 ⓒ. In the same way, the results for experts are presented. Figure 5.5 ⓓ and ⓔ visualize the number of errors with respect to each expert and the respective task while the median values for each task are shown in Figure 5.5 ⓕ.

Furthermore, the data has to be checked for its normal distribution since this is necessary for further analysis (cf. Section 5.3). As an example, Figure 5.6 graphically presents the distribution of the data (total time) regarding the group of experts (task 1) while Table 5.3 shows the results of two different statistical tests of normality (Shapiro–Wilk test and Anderson–Darling test [18, 19]).

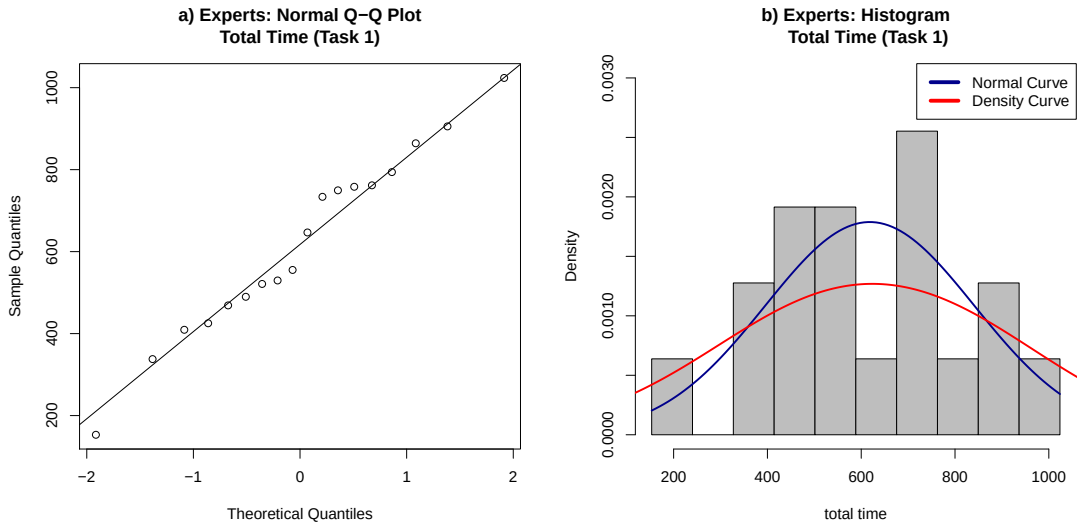


Figure 5.6: Experts: Distribution of Total Time (Task 1)

In particular, Figure 5.6 ⓐ shows a quantile-quantile plot (Q-Q plot) plotting the quantiles of the sample distribution against the quantiles of a theoretical distribution (i.e., normal distribution). If the points follow the line, one can suggest normal distribution of the data [20]. Figure 5.6 ⓑ shows a histogram presenting the probability densities. Furthermore, the normal distribution curve (blue line) as well as the density curve (red line) are plotted

within this graphic in order to get a better understanding of the data distribution. Table 5.3 shows the results (p-values) of two statistical tests of normality. Therefore, the statistic software R [20] was used to calculate the respective p-values. The latter represent probabilities ($0 \leq p \leq 1$) and can be used for hypothesis testing. A p-value above a specific significance level (α, mostly 0.05) indicates a non-significant result and, therefore, the null hypothesis can not be rejected. On the contrary, if the p-value is less (or equal) α, the null hypothesis can be rejected. Both statistical tests used assume a normal distribution of the data within their null hypotheses while their alternative hypotheses state that there is no normal distribution.

Task	**Shapiro–Wilk test**	**Anderson–Darling test**
Task 1	0.945432	0.765591

Table 5.3: Experts: P-Values Total Time (Task 1)

Considering the results of Table 5.3 as well as the graphical representation (cf. Figure 5.6) we can assume that the data of the total time for the group of experts is normally distributed. In particular, the p-values of the statistical tests, as mentioned above (0.945432 and 0.765591), are greater than the significance level α of 0.05 and, therefore, the null hypotheses can not be rejected.
More plots and test results concerning the normal distribution of the remaining data sets are presented in Appendix F.

Additionally, results regarding the mental effort and comprehension questionnaires are presented in the following.
Figure 5.7 visualizes the results regarding the mental effort questionnaire the subjects had to deal with after each task (cf. Table 3.4). Figure 5.7 ⓐ and ⓑ present the median values regarding the group of novices separated by task while Figure 5.7 ⓒ and ⓓ accordingly show the results (median) for the group of experts.
The median values of the results for the total time as well as for the number of errors regarding the use of screencasts are presented in Figure 5.8. As mentioned before, the subjects were able to take a look at screencasts during modeling the tasks if needed (cf. Section 3.6).

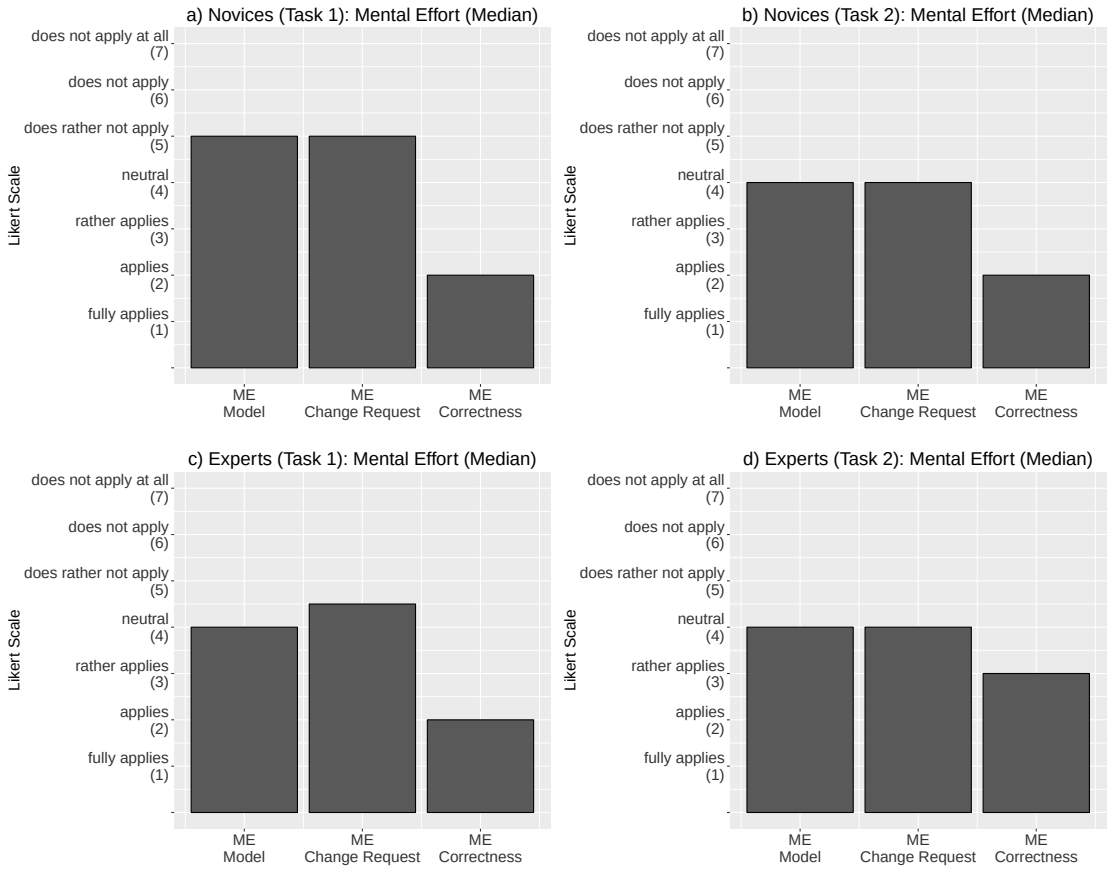


Figure 5.7: Mental Effort (ME) per Task (Median)

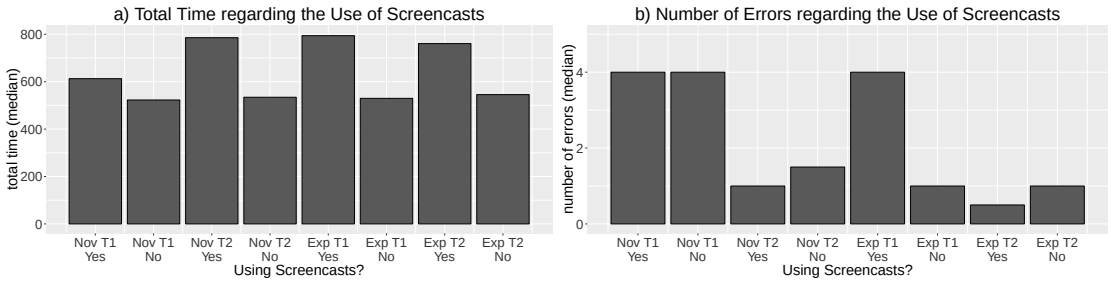


Figure 5.8: Summary Use of Screencasts (Median)

After each task, the subjects were asked if they needed to take look at the screencasts (cf. Table 3.4). Considering those answers Figure 5.8 ⓐ presents median values of the results for the total time subjects needed to solve each task separated by the different groups (novices as nov, experts as exp), as well as whether or not they have watched the screencasts. Generally, one can observe that subjects who have to take a look at the screencasts need more time solving the tasks. In the same way the results regarding the number of errors are presented in Figure 5.8 ⓑ.

Furthermore, Figure 5.9 illustrates the results referring to the mental effort questionnaire the subjects have to deal with at the end of the experiment (cf. Table 3.5). Therefore, Figure 5.9 ⓐ presents the median values of the results with respect to the group of novices while Figure 5.9 ⓑ shows those for the group of experts.

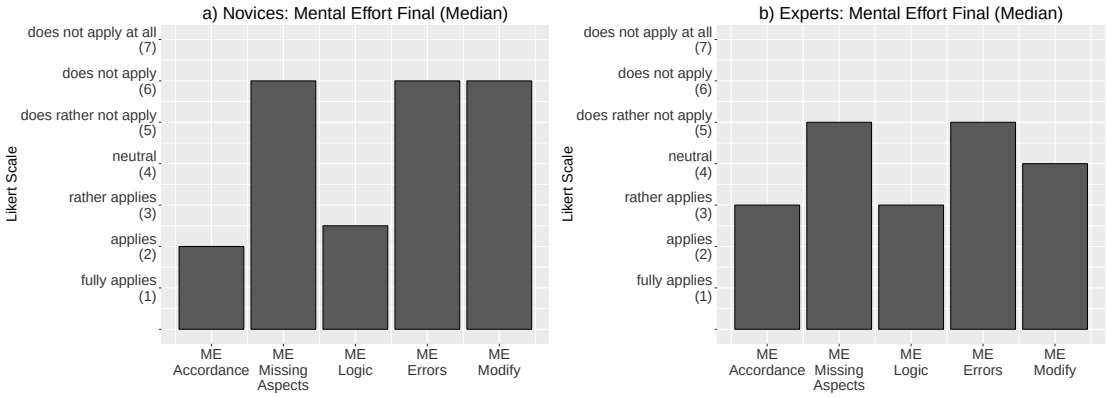


Figure 5.9: Mental Effort (ME) Final Questionnaire (Median)

Group	Achieved Score (Median)	Possible Score
Novices	20.5	25
Experts	21.5	25

Table 5.4: Comprehension Questionnaire Results

The results of the comprehension questionnaire (cf. Section 3.6 and Appendix C) are presented in Table 5.4. One can observe that novices gain less points (median of 20.5 points) solving this questionnaire than experts (median of 21.5 points). The possible score to be achieved was 25 points. Generally, the points were calculated as follows:

Questions 1-10 were single choice or single yes / no questions. Therefore, a correct answer resulted in 1 point, otherwise 0 points. Question 11 and 12 were multiple choice questions and the points were calculated by summing up the correct answer options (cao) as well as the false answer options (fao). Finally, the achieved points for this type of question were calculated by subtracting fao from cao (possible negative values were treated as 0 points):

$$points = \max(0, cao - fao)$$

Summarizing the points of each question resulted in the final achieved score.

Finally, we derived further variables (i.e., number of activities, change request time (cf. Figure 5.10)) from the data obtained by the log files (cf. Table 3.2). In particular, the

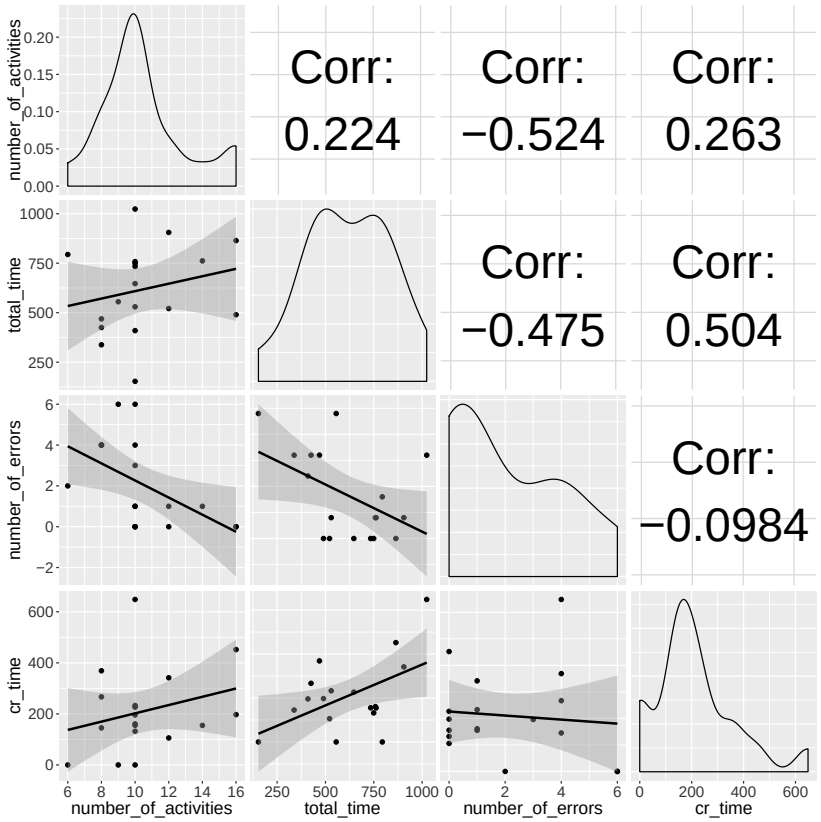


Figure 5.10: Matrix Plot Experts Task 1

number of activities specifies the amount of operations a subject needed in order to complete a specific task (represented as row in the log file, cf. Table 3.2). The change request time (cr_time) represents the time a subject needed to perform the explicit change request given in each task.
The matrix plot in Figure 5.10 illustrates an example for the group of experts regarding task 1. Therefore, scatter plots show the different dependencies between the additionally calculated variables as well as the variables already introduced in this section (i.e., total time, number of errors). Beside the scatter plots, on the diagonal, density curves are visualized representing the distribution of the respective data set. The upper right section shows the Pearson product-moment correlation coefficient (Pearson's r) specifying the linear dependence between the respective variables. The correlation coefficient is a single value between -1 and 1 describing the relationship between two variables. -1 means there exits a negative relationship (high values on one variable are connected with low values on the other), while 1 means there exists a positive relationship (high values on one variable are connected with high values on the other). Furthermore, a value of 0 indicates no relationship between the two variables [12, 20]. The results regarding the other groups and tasks are shown in Appendix G.

Further interpretation of the results introduced in this section are discussed in Section 5.4.

5.2 Data Set Reduction

When applying statistical methods during the analysis phase, it is important to ensure that all data is valid in order to draw correct conclusions. Therefore, the quality of input data has to be checked accordingly before processing. As mentioned and seen before (cf. Figure 5.2 ⓐ), anomalies in data may occur in the form of outliers. After identifying the latter (e.g., using box plots), it is important to decide how to handle them with respect to the information they contain. Outliers are only candidates for removal when dealing with data set reduction. However, this does not mean that they have to be removed

in any case. When removing an outlier it is important to guarantee that no significant important information is lost within the modified results [4].

In our experiment, several outliers can be observed as shown in the Figures 5.2 and 5.3 (in each case ⓐ and ⓒ). Two examples are:

- One subject needed 1441.05 seconds for modeling task 2
- One subject made 15 errors in the resulting questionnaire model of task 2

After considering all outliers carefully, it was decided that the results seem to be valid for the experiment and, therefore, to not remove them in order to avoid a loss of information.

5.3 Hypothesis Testing

This section introduces hypothesis testing and compares the hypotheses from Section 3.3 with the findings in the data collected. After the data has been evaluated by descriptive statistics (cf. Section 5.1) further analysis has to be done including statistical methods in order to investigate the stated hypotheses. Descriptive statistics may present differences and, therefore, assumptions regarding the hypotheses can be made but have to be proofed. Therefore, [4] introduces different statistical methods in order to test the hypotheses. The intention of hypothesis testing is to reject the null hypothesis with a significance as high as possible. As mentioned in Section 5.1, p-values can be calculated in order to decide whether a result of a specific test is significant or not (in combination with a significance level α). Based on the significance of the result, further decisions about the null hypothesis can be made. In particular, a significant result (p-value $\leq \alpha$ (0.05)) leads to a rejection of the null hypothesis and, therefore, to an acceptance of the alternative hypothesis. On the contrary, the null hypothesis can not be rejected if a non-significant result (p-value $> \alpha$ (0.05)) is observed. However, this does not proof a failure of the alternative hypothesis [4, 12].

Considering the results of the normality tests (cf. Section 5.1 and Appendix F), several statistical tests have been conducted in order to test the hypotheses described in Section

3.3. Again, the statistic software R [20] was used to calculate the respective p-values and the results are shown in Table 5.5.

Hypothesis $H1$:			
Test	**Samples Involved**	**p-value**	**Significant?**
t-test	toti_nt1 / toti_et1	0.8684	no
t-test	toti_nt2 / toti_et2	0.3668	no
t-test	toti_nt1_nt2 / toti_et1_et2	0.1089	no
Hypothesis $H2$:			
Test	**Samples Involved**	**p-value**	**Significant?**
paired t-test	toti_nt1 / toti_nt2	0.04606	yes
Hypothesis $H3$:			
Test	**Samples Involved**	**p-value**	**Significant?**
paired t-test	toti_et1 / toti_et2	0.4275	no
Hypothesis $H4$:			
Test	**Samples Involved**	**p-value**	**Significant?**
u-test	noe_nt1 / noe_et1	0.08438	no
u-test	noe_nt2 / noe_et2	0.2737	no
u-test	noe_nt1_nt2 / noe_et1_et2	0.7002	no
Hypothesis $H5$:			
Test	**Samples Involved**	**p-value**	**Significant?**
Wilcoxon test	noe_nt1 / noe_nt2	0.7884	no
Hypothesis $H6$:			
Test	**Samples Involved**	**p-value**	**Significant?**
Wilcoxon test	noe_et1 / noe_et2	0.2304	no

Table 5.5: Results of Hypothesis Testing

The first hypothesis ($H1$) deals with the total times the subjects needed to solve the given tasks considering the different experience. Since the data sets for the total time of both groups are normally distributed as well as independent, we decided to use the parametric *One-Tailed T-Test* as described in [4, 12]. Therefore, three different t-tests have been performed. The first one compares the data set of total time for the group of novices (toti_nt1) with the one for the group of experts regarding task 1 (toti_et1). The

second test compares the total times of both groups regarding task 2 (toti_nt2, toti_et2) while the third test compares the groups regarding the time differences between task 1 and task 2 (toti_nt1_nt2, toti_et1_et2). The single time differences are calculated by subtracting the times of task 1 from those of task 2.

The second hypothesis ($H2$) deals with the total times novices needed to solve the given tasks regarding their different difficulties. Therefore, the total times of the two tasks (toti_nt1, toti_nt2) within this group have to be compared. Since the data sets are normally distributed but not independent we decided to use the *One-Tailed Paired T-Test* in order to test this hypothesis [4, 12].

The same type of test has been applied to test the third hypothesis ($H3$). Therefore, the total times of the two tasks (toti_et1, toti_et2) within the group of experts have to be compared.

The fourth hypothesis ($H4$) deals with the number of errors the subjects made in the resulting questionnaires considering their different experience in process modeling. Since the respective data sets are not normally distributed but independent, we decided to apply the non-parametric *One-Tailed Wilcoxon-Mann-Whitney Test* (u-test) as described in [4, 20]. Again, three different tests have been performed. The first one deals with the comparison of both groups regarding the number of errors made in task 1 (noe_nt1,noe_et1). The second one compares the number of errors regarding task 2 (noe_nt2,noe_et2), while the third one compares the groups regarding the differences in errors between task 1 and task 2 (noe_nt1_nt2, noe_et1_et2). Again, the single differences are calculated by subtracting the number of errors made in task 1 from those made in task 2.

The fifth hypothesis ($H5$) deals with the number of errors novices made in the resulting questionnaire models regarding the different difficulties of the tasks. Therefore, the number of errors made in each task have to be compared within this group (noe_nt1, noe_nt2). Since the data sets are not normally distributed but dependent, we decided to use the non-parametric *One-Tailed Wilcoxon Test*. The latter is an alternative to the paired t-test [4].

In the same way the sixth hypothesis ($H6$) has been tested. The latter deals with the number of errors experts made in the resulting questionnaire models regarding the

different difficulties of the tasks. Therefore, the two tasks have been compared (noe_et1, noe_et2).
Considering the results in Table 5.5, only the test for **H2** shows a significant result (p-value = 0.04606) and, therefore, the null hypothesis can be rejected and the alternative hypothesis can be accepted. All other tests show non-significant results (p-value > 0.05). Therefore, the null hypotheses of **H1**, **H3**, **H4**, **H5** as well as **H6** can not be rejected and need to be accepted.
The following section introduces a summary of the experiment analysis as well as an interpretation and discussion of the obtained results.

5.4 Summary and Discussion

This section introduces an interpretation of the results obtained regarding the objective of the study. The latter was to investigate how much effort domain experts need in order to properly handle a newly developed software application (Questioneer, introduced in Section 3.1). In particular, the following fundamental research question has to be answered:

Do end-users (e.g., medical doctors, psychologists) understand the (modeling) concept of the software application Questioneer, with respect to the complexity of the provided application?

To answer this question, an experiment was conducted which involved students and research associates with different background knowledge in process modeling. The aim was to investigate whether or not special requirements (i.e., prior knowledge in process modeling) are needed to properly handle Questioneer. Therefore, the subjects were separated into two groups (novices and experts (cf. Section 3.4)) in order to investigate if there are differences in effort needed (i.e., time and number of errors) between those groups when modeling given tasks with Questioneer. Furthermore, possible differences have to be investigated when modeling tasks with increasing complexity regarding the subject's prior experience in process modeling.

Considering the results of hypothesis testing (cf. Section 5.3) only one significant result was obtained ($H2$) when testing several aspects. Thus, the alternative hypothesis that novices are significantly slower in solving tasks with higher difficulty is the only assumption that can be accepted. Surprisingly, all other assumptions (stated in the alternative hypotheses described in Section 3.3) can not be accepted. In particular, the assumption that experts are faster with respect to solving the required tasks than novices ($H1$) could not be proofed. Considering the results in Figure 5.4 and Table 5.1 we can even observe that novices especially needed less time when solving task 1 than experts. However, in consideration of the results regarding the number of errors (cf. Table 5.2), novices made more errors in the resulting questionnaire model of task 1. This may be explained by the fact, that end-users with less or no prior experience in process modeling are not as conscientious as end-users with experience. Novices possibly do not focus on details needed to model the questionnaire properly. The assumption, that experts are significantly slower in solving tasks with higher difficulty ($H3$), can not be accepted regarding the results of hypothesis testing. Although, we can observe a little increase of the time (cf. Table 5.1) the differences are not significant. Thus, we must assume that a higher complexity of questionnaire models, which have to be developed within Questioneer, does not significantly affect the time an end-user with prior experience in process modeling needs. Furthermore, the expectation that experts make less errors than novices ($H4$) could not be proofed with this study. Although, regarding task 1, we can observe a difference (novices made more errors than experts, (cf. Table 5.2)) the results, however are not significant enough in order to reject the null hypothesis. Therefore, we must assume that there are no significant differences in the number of errors made regarding the user's experience in process modeling. Another assumption that could not be proofed with the data collected, is that novices make significant more errors when solving tasks with higher difficulty ($H5$). Considering the results of Figure 5.5 and Table 5.2 we can even observe that novices made less errors when increasing the complexity of the model. It is conceivable, that a kind of learning effect has taken place during the experiment regarding users with less or no prior experience in process modeling. In particular, novices possibly gained knowledge while modeling the first task and, therefore, made less errors when modeling the second task.

The last assumption, that users with experience in process modeling make significant more errors when solving tasks with higher difficulty ($H6$), could neither be proofed. The results, in turn, show that there are no significant differences in the number of errors made when increasing the complexity of the models. Finally, no significant differences can be observed regarding the results of the comprehension questionnaire (cf. Table 5.4).

Summarizing the mentioned aspects, the experiment showed that there are no significant differences in handling the software application regarding the user's experience in process modeling. Although, that does not proof that there are no differences, a first assumption can be made regarding the overall understanding of the (modeling) concept used within the questionnaire configurator Questioneer. The assumption is that prior knowledge in process modeling is not necessarily required in order to understand and handle the software application properly. The experiment even showed, as mentioned before, that users with less or no prior experience may gain enough knowledge within one hour (the duration of the experiment conducted) in order to operate as properly as users with experience. Therefore, we may attribute a certain intuitiveness of the modeling concept to Questioneer. Regarding the number of errors, we can observe that, in general, only few errors were made (cf. Table 5.2). In particular, 8 novices (out of 24) and 10 experts (out of 18) made less (or equal) 1 error regarding task 1 while 13 novices and 12 experts made less (or equal) 1 error regarding task 2. That, in turn, may be another argument for a certain intuitiveness of the modeling concept.

Although, the experiment provides first indications of an intuitive modeling concept of the questionnaire configurator additional experiments (i.e., replication with more subjects) are recommended in order to strengthen the assumption. Another possibility to proof the usability of the application may be further experimentation considering different focal points with respect to the subjects. As an example, users with no experience in using Questioneer may be compared with users that have already used Questioneer for a longer period of time.

6

Related Work

This thesis evaluates a newly developed configurator application for modeling data collection instruments. More precisely, the overall understanding of its (modeling) concept is investigated. Therefore, our work is related to the following aspects: total time needed to solve specific tasks within this application as well as the number of errors made in the resulting questionnaire models. Based on these aspects, conclusions about the intuitiveness of the configurator application are drawn.

By now, little work exists considering the topic of empirical evaluation of software applications. Even in the whole area of software engineering there exists a lack of experimental validation [21]. However, [22] introduces a web-based configurator for context-aware experience sampling apps in ambulatory assessment (*ESMAC*). Therefore, the system is evaluated by testing its web-interface as well as its Android application against the *movisensXS*[4] platform. For investigations, two studies are conducted. The first study (regarding the web-interface of the configurator) involved two ambulatory assessment experts designing an experience sampling methods (ESM) study by using each platform (ESMAC and movisensXS). The second study (regarding the Android application) involved 10 subjects divided into two groups. Both groups had to work with each application for three days. In contrary to our experiment, both studies use the *System Usability Scale (SUS)* described in [23] in order to measure usability.

Another web-application *SCAMP* (Search ConfigurAtor for experiMenting with PuppyIR) is introduced and investigated in [24]. SCAMP is described as a web-based application in order to create, conduct, coordinate as well as monitor interactive information retrieval (IIR) experiments. The configurator application is evaluated in two ways. On

[4] https://xs.movisens.com/ (last accessed on 2016/11/11)

the one hand, an usability analysis is performed by an human-computer interaction (HCI) researcher as well as an undergraduate student by setting up an IIR experiment. Afterwards, they state their opinion about the usability. On the other hand, an IIR experiment is set up within this application and 48 participants were asked to perform the experiment. Besides the several monitoring features provided by SCAMP itself (i.e., timings) the participants were asked afterwards how they believe how well they had performed. Nonetheless, both studies focus on different aspects as our work in order to gain information about the usability of the respective application.

A controlled experiment for empirical evaluating a tool called *CodeCity* (a 3D software visualization approach) is introduced in [25]. The experiment investigates the effectiveness as well as efficiency of the approach in comparison to state-of-the-practice approaches. The experiment focuses on the correctness as well as on the completion time of the task 45 participants had to process. In contrary to our experiment, this study evaluates a software application compared to one that is state-of-the-art. Considering a configurator application for modeling data collection instruments, no similar approach is available which can be compared with Questioneer.

Although, all these works focus in some way on the usability of the respective object, our experiment pursues a different approach. In particular, the experiment conducted in this thesis investigated the configurator application by observing correctness aspects and completion time when solving specific tasks. Thereby, no comparison to another application has taken place. We argue, that using these metrics intuitiveness of the (modeling) concept of the application can be shown as well. Especially, participants with less or no prior experience in process modeling show a good learning curve regarding the errors made when solving given tasks.

7

Conclusion

This thesis investigated the questionnaire configurator Questioneer with respect to its usability. In particular, a controlled experiment with 44 participants was conducted to get further insights about the overall understanding of the (modeling) concept with respect to the complexity of the application. For this purpose, the participants were separated into two groups based on their background knowledge and experience in process modeling. During the experiment, each group had to create two questionnaires (each with different complexity) only using the provided configurator application. In order to determine possible differences between those groups, we focused on the total time the subjects needed to solve a specific task as well as the number of errors they made in the resulting questionnaire models. Furthermore, possible differences have been investigated when modeling tasks with increasing complexity regarding the subject's prior experience in process modeling.

Even if the data obtained during the experiment, in several aspects, shows differences regarding the total times as well as the number of errors, the results are not significant enough. Therefore, the latter are not in accordance with the defined alternative hypotheses. Contrary to the expectations and assumptions (except the assumption of the second alternative hypothesis) stated within the alternative hypotheses, the experiment showed that there are no significant differences in handling the software application regarding the user's experience in process modeling.

Due to these results, a final assumption can be made regarding the usability of the questionnaire configurator. Considering the user's experience in process modelling there is no need for prior knowledge in order to understand and handle the software application properly. Furthermore, a very interesting aspect showed with the experiment

is that users with less (or no) prior experience in process modeling may gain knowledge in an adequate time in order to properly handle the application. We consider this as some kind of learning effect. Therefore, one can assume that the modeling concept of the software application Questioneer is intuitive and that end-users are able to create questionnaires adequately with respect to the complexity of the application.
Nonetheless, additional research is highly recommended in order to strengthen the assumption about the intuitiveness of the modeling concept of the questionnaire configurator Questioneer. Moreover, further experiments (i.e., different environments, different groups of participants) are necessary in order to obtain more results allowing to confirm the outcome.

Bibliography

[1] Basili, V.R.: The role of experimentation in software engineering: Past, current, and future. In: Proceedings of the 18th International Conference on Software Engineering. ICSE '96, Washington, DC, USA, IEEE Computer Society (1996) 442–449

[2] ISO: IEC25010: 2011 Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models (2011)

[3] Rafique, I., Lew, P., Abbasi, M.Q., Li, Z.: Information Quality Evaluation Framework: Extending ISO 25012 Data Quality Model. World Academy of Science, Engineering and Technology **65** (2012) 523–528

[4] Wohlin, C., Runeson, P., Höst, M., Ohlsson, M.C., Regnell, B., Wesslén, A.: Experimentation in Software Engineering - An Introduction. Kluwer Academic Publisher (2000)

[5] Schobel, J., Pryss, R., Schickler, M., Ruf-Leuschner, M., Elbert, T., Reichert, M.: End-User Programming of Mobile Services: Empowering Domain Experts to Implement Mobile Data Collection Applications. In: IEEE 5th Int'l Conf on Mobile Services, IEEE Computer Society Press (2016)

[6] Schobel, J., Pryss, R., Schickler, M., Reichert, M.: Towards Flexible Mobile Data Collection in Healthcare. In: 29th IEEE International Symposium on Computer-Based Medical Systems (CBMS 2016). (2016) 181–182

[7] DBIS Ulm: QuestionSys - A Generic and Flexible Questionnaire System Enabling Process-Driven Mobile Data Collection. Internet WWW page, at URL: https://www.uni-ulm.de/in/iui-dbis/forschung/laufende-projekte/questionsys/ (2016) , last accessed on 2016/11/11.

[8] Schobel, J., Pryss, R., Schickler, M., Reichert, M.: A Configurator Component for End-User Defined Mobile Data Collection Processes. In: ICSOC'16, Demo Track of the 14th Int'l Conf on Service Oriented Computing. LNCS, Springer (2016)

[9] Basili, V.R., Selby, R.W., Hutchens, D.H.: Experimentation in software engineering. IEEE Transactions on Software Engineering **SE-12** (1986) 733–743

[10] Munassar, N.M.A., Govardhan, A.: A Comparison Between Five Models Of Software Engineering. IJCSI **5** (2010) 94–101

[11] Basili, V.R.: Software Modeling and Measurement: The Goal/Question/Metric Paradigm. Technical report, College Park, MD, USA (1992)

[12] Trochim, W.M.: The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: http://www.socialresearchmethods.net/kb/ (version current as of October 20, 2006) , last accessed on 2016/11/11.

[13] Pfleeger, S.L.: Experimental design and analysis in software engineering. Annals of Software Engineering **1** (1995) 219–253

[14] Zimoch, M.: Influence of Psychological Distance on Process Modeling: A Gamification Approach. Master's thesis, Ulm University (2015)

[15] Cook, T.D., Campbell, D.T.: Quasi-Experimentation: Design & Analysis Issues for Field Settings. Houghton Mifflin (1979)

[16] Michael Frigge, David C. Hoaglin, B.I.: Some Implementations of the Boxplot. The American Statistician **43** (1989) 50–54

[17] Upton, G., Cook, I.: Understanding Statistics. Oxford University Press (1996)

[18] Shapiro, S.S., Wilk, M.B.: An analysis of variance test for normality (complete samples). Biometrika **52** (1965) 591–611

[19] Razali, N.M., Wah, Y.B.: Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. Journal of Statistical Modeling and Analytics **2** (2011) 21–33

[20] Team, R.C.: The R Project for Statistical Computing. Internet WWW page, at URL: https://www.r-project.org/ (2016) , last accessed on 2016/11/11.

[21] Tichy, W.F., Lukowicz, P., Prechelt, L., Heinz, E.A.: Experimental Evaluation in Computer Science: A Quantitative Study. Journal of Systems and Software **28** (1995) 9–18

[22] Bachmann, A., Zetzsche, R., Schankin, A., Riedel, T., Beigl, M., Reichert, M., Santangelo, P., Ebner-Priemer, U.: ESMAC: A Web-Based Configurator for Context-Aware Experience Sampling Apps in Ambulatory Assessment. In: Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare. MOBIHEALTH'15, ICST, Brussels, Belgium, Belgium, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering) (2015) 15–18

[23] Brooke, J., et al.: SUS: a 'quick and dirty' usability scale. Usability Evaluation in Industry **189** (1996) 4–7

[24] Renaud, G., Azzopardi, L.: SCAMP: A Tool for Conducting Interactive Information Retrieval Experiments. In: IIiX. (2012) 286–289

[25] Wettel, R., Lanza, M., Robbes, R.: Software Systems as Cities: A Controlled Experiment. In: Proceedings of the 33rd International Conference on Software Engineering, ACM (2011) 551–560

A

Error Evaluation Sheet

Figure A.1 presents the different types of errors, their classification as well as their weighting.

Error Evaluation Sheet

Occurring Errors:	Weighting (W):	Amount of Errors (A):	Error Points (W * A):
Regarding Complex Elements (Pages):			
- wrong order of the elements:	1	_____	_____
- missing element:	2	_____	_____
- too many elements (wrong element)	1	_____	_____
- element in wrong page	1	_____	_____
- missing page (additional to missing element)	1	_____	_____
- wrong page order:	1	_____	_____
Regarding Simple Elements:			
- missing item (e.g., Ausstattung)	2	_____	_____
Regarding Desicions (subsequent faults):			
- missing decision (additional to following errors)	1	_____	_____
- missing answer option	1	_____	_____
- missing path	2	_____	_____
- missing condition	2	_____	_____
- page on wrong path	1	_____	_____
		Total Error Points:	_____

Figure A.1: Error Evaluation Sheet (per Task)

B

Task Sheets

ulm university universität ulm

Fakultät für
Ingenieurwissenschaften, Informatik und Psychologie
Datenbanken und Informationssysteme

Experiment: Questioneer

Aufgabe 1 ID-Nr:2173

Vorwort:

In dieser Aufgabe sollen Sie einen Fragebogen modellieren, der es einem Benutzer ermöglicht einen passenden PC zu konfigurieren.

Lesen Sie zunächst die Aufgabe einmal komplett und aufmerksam durch!
Questioneer soll dabei geschlossen bleiben!

Folgender Fragebogen soll mit Hilfe von Questioneer erstellt werden:

1. Als erste Seite soll dem Benutzer eine Willkommen-Seite angezeigt werden.
   Diese beinhaltet den Namen des Fragebogens als Überschrift und danach einen kurzen Einleitungstext.

---

2. Anschließend soll dem Benutzer eine Seite angezeigt werden, auf der er Angaben bzgl. des Prozessors in seinem neuen PC machen kann.
   Diese beinhaltet der Reihe nach einen kleinen Text, dann die Frage nach dem Prozessor (AMD oder Intel) und anschließend die Frage wie viele CPU-Kerne dieser besitzen soll.

---

3. Auf einer nächsten Seite soll der Benutzer Angaben bzgl. der Ausstattung seines neuen PCs machen können.
   Der Reihe nach beinhaltet diese einen kleinen Text sowie anschließend die Frage nach der Ausstattung selber.

---

4. Als nächstes soll der Benutzer Angaben zu dem Verwendungszweck machen können.
   Beginnend mit einem kleinen Text soll er danach nach seinem primären Verwendungszweck (Office-Anwendungen, Web-Anwendungen, Gaming) gefragt werden und anschließend Angaben darüber machen, wie sehr er darauf den Fokus legt.

Figure B.1: Task 1 - Part 1

5. Um mit dem Benutzer in Kontakt treten zu können, sollen dessen Kontaktdaten auf einer nächsten Seite erfasst werden.
   Nach einer Überschrift und einem kurzen Einleitungstext, soll er nach Namen, Vornamen, Telefon, Adresse und E-Mail gefragt werden.

---

6. Abschließend soll dem Benutzer ein Seite angezeigt werden, um sich bei ihm zu bedanken.
   Diese beinhaltet eine Überschrift und anschließend einen Text mit der Danksagung.

---

7. Speichern Sie Ihr Modell mittels "rechtsklick" → "save image" auf dem Desktop.
   Verwenden Sie als Dateiname bitte 2173_Aufgabe1_v1.png!

---

8. Das Modell soll nun abgeändert werden.
   In der Seite von Punkt 3 soll zusätzlich das Gehäuse des PCs erfragt werden.
   Fügen Sie die vordefinierte Frage "Gehäuse" an das Ende der Seite aus Punkt 3 ein, und passen sie ihr schon erstelltes Modell dementsprechend an.

---

9. Speichern Sie das neue Modell mittels "rechtsklick" → "save image" auf dem Desktop.
   Verwenden Sie als Dateiname bitte 2173_Aufgabe1_v2.png!

Nachdem Sie die Aufgabe durchgelesen haben **starten** Sie Questioneer, indem Sie "questioneer.exe" im Ordner "studie" auf Ihrem Desktop ausführen.

Bitte wählen Sie für diese Aufgabe nun folgende Arbeitsumgebung aus:

Workspace: **Experiment**
Questionnaire: **Aufgabe1**

Navigieren Sie bitte anschließend in den Bereich "Editor", dort in den **dritten** Bereich (Schaltfläche **3**) und beginnen Sie mit dem Modellieren des oben beschriebenen Fragebogens.

Sofern vorhanden, dürfen Sie die **vordefinierten Elemente und Seiten** verwenden.

Im Ordner "studie" finden Sie zusätzlich einen Ordner "Screencasts". Darin finden Sie verschiedene Screencasts bzgl. der Bedienung von Questioneer. Sollten Sie Probleme beim Lösen der Aufgabe haben, können Sie sich den entsprechenden Screencast ansehen, indem Sie die ".htm-Datei" öffnen.

Nachdem Sie der Meinung sind die Aufgabe erledigt zu haben und Sie Ihr Modell gespeichert haben BEENDEN Sie Questioneer!

Beachten Sie bitte anschließend die Rückseite dieses Aufgabenblattes.

Figure B.2: Task 1 - Part 2

ulm university universität ulm

Fakultät für
Ingenieurwissenschaften, Informatik und Psychologie
Datenbanken und Informationssysteme

Experiment: Questioneer

Aufgabe 2 ID-Nr:2173

Vorwort:

In dieser Aufgabe sollen Sie nochmals einen Fragebogen modellieren, der es einem Benutzer ermöglicht einen passenden PC zu konfigurieren.

Lesen Sie zunächst die Aufgabe einmal komplett und aufmerksam durch!
Questioneer soll dabei geschlossen bleiben!

Folgender Fragebogen soll mit Hilfe von Questioneer erstellt werden:

1. Als erste Seite soll dem Benutzer eine Willkommen-Seite angezeigt werden.
   Diese beinhaltet den Namen des Fragebogens als Überschrift und danach einen kurzen Einleitungstext.

---

2. Auf einer nächsten Seite soll der Benutzer Angaben bzgl. der Ausstattung seines neuen PCs machen können.
   Der Reihe nach beinhaltet diese einen kleinen Text sowie anschließend die Frage nach der Ausstattung selber.
   Dazu soll die Frage "Ausstattung" zuerst abgeändert werden.
   Fügen Sie der Frage "Ausstattung" eine weitere Antwortmöglichkeit (Item) "Betriebssystem" hinzu.

---

3. Basierend auf der Information über die Ausstattung soll eine weitere Seite angezeigt werden.
   Hat der Benutzer bei seiner Ausstattung das Betriebssystem gewählt, soll ihm eine Seite angezeigt werden, die eine Abfrage über das jeweilige Betriebssystem erlaubt.
   Diese Seite soll zuerst einen kurzen Text und anschließend die entsprechende Abfrage, bzgl. welches Betriebssystem installiert sein soll, enthalten.

---

4. Unabhängig von der Ausstattung soll es dem Benutzer in einer weiteren Seite möglich sein, Angaben zu optionalen Komponenten zu machen.
   Diese beinhaltet der Reihe nach einen kurzen Text, die Frage nach einem Monitor und abschließend die Frage nach einem Drucker.

Figure B.3: Task 2 - Part 1

5. Anschließend sollen weitere Informationen erfragt werden.
   Auf dieser Seite werden nach einer kurzen Überschrift, die Fragen nach dem aktuellen Gerät, dem geplanten Budget und dem Zeitpunkt des Kaufes (in dieser Reihenfolge) gestellt.

---

6. Um mit dem Benutzer in Kontakt treten zu können, sollen dessen Kontaktdaten auf einer nächsten Seite erfasst werden.
   Nach einer Überschrift und einem kurzen Einleitungstext, soll er nach Namen, Vornamen, Telefon, Adresse und E-Mail gefragt werden.

---

7. Abschließend soll dem Benutzer ein Seite angezeigt werden, um sich bei ihm zu bedanken.
   Diese beinhaltet eine Überschrift und anschließend einen Text mit der Danksagung.

---

8. Speichern Sie Ihr Modell mittels "rechtsklick" → "save image" auf dem Desktop.
   Verwenden Sie als Dateiname bitte 2173_Aufgabe2.png!

Nachdem Sie die Aufgabe durchgelesen haben **starten** Sie Questioneer, indem Sie "questioneer.exe" im Ordner "studie" auf Ihrem Desktop ausführen.

Bitte wählen Sie für diese Aufgabe nun folgende Arbeitsumgebung aus:

Workspace: **Experiment**
Questionnaire: **Aufgabe2**

Navigieren Sie bitte anschließend in den Bereich "Editor", dort in den **dritten** Bereich (Schaltfläche **3**) und beginnen Sie mit dem Modellieren des oben beschriebenen Fragebogens.

Sofern vorhanden, dürfen Sie die **vordefinierten Elemente und Seiten** verwenden.

Im Ordner "studie" finden Sie zusätzlich einen Ordner "Screencasts". Darin finden Sie verschiedene Screencasts bzgl. der Bedienung von Questioneer. Sollten Sie Probleme beim Lösen der Aufgabe haben, können Sie sich den entsprechenden Screencast ansehen, indem Sie die ".htm-Datei" öffnen.

Nachdem Sie der Meinung sind die Aufgabe erledigt zu haben und Sie Ihr Modell gespeichert haben BEENDEN Sie Questioneer!

Beachten Sie bitte anschließend die Rückseite dieses Aufgabenblattes.

Figure B.4: Task 2 - Part 2

C

Comprehension Questionnaire

ulm university universität ulm

Fakultät für
Ingenieurwissenschaften, Informatik und Psychologie
Datenbanken und Informationssysteme

Experiment: Questioneer

Verständnisfragen ID-Nr:2173

Beantworten Sie folgende Fragen zu "Questioneer" **ohne** dabei die Screencasts anzusehen!

1. Kann man verschiedene Workspaces auswählen?

   ◯ ja ◯ nein

2. Kann man verschiedene Fragebögen auswählen?

   ◯ ja ◯ nein

3. Kann "Questioneer" verschiedene Sprachen verwalten?

   ◯ ja ◯ nein

4. Was bedeutet die markierte Ziffer im Bild?
   Bitte wählen Sie **eine** der folgenden Antworten.

   ○ Die Anzahl an Elementen im Fragebogen
   ○ Die Revisionsnummer dieser Seite
   ○ Die Anzahl an Elementen in einer Seite
   ○ Die Anzahl an Seiten im Fragebogen
   ○ Die Anzahl an Versionen dieser Seite

1

Figure C.1: Comprehension Questionnaire - Part 1

5. Wie kommt es zu einer Erhöhung der Revisionsnummer?
   - ○ Die Revisionsnummer wird automatisch beim Editieren eines Elements bzw. einer Seite erhöht
   - ○ Die Edition muss explizit als neue Revision gespeichert werden

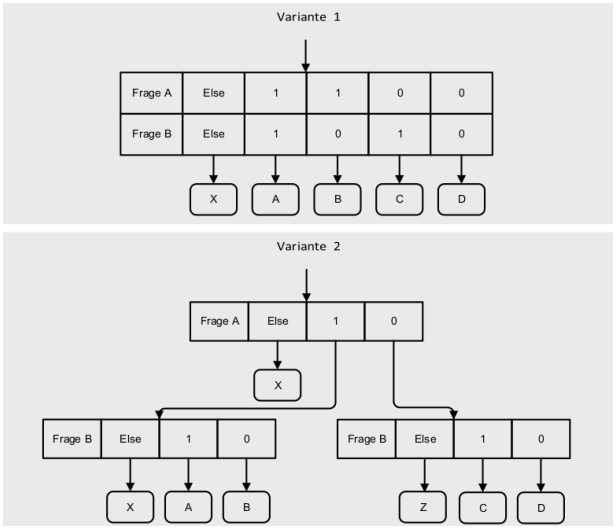
6. Ist es möglich Seiten aus dem Fragebogen zu entfernen?
   - ○ ja ○ nein

7. Ist es möglich Entscheidungen (Decisions) in einen Fragebogen zu integrieren?
   - ○ ja ○ nein

8. Auf welche Fragen kann sich eine Entscheidung im Fragebogen-Modell beziehen?
   - ○ Ausschließlich auf Fragen aus der vorhergehenden Seite
   - ○ Auf alle Fragen die vor dieser Entscheidung liegen

9. Wie viele Fragen können in eine einzige Entscheidung integriert werden?
   - ○ beliebig viele ○ genau eine

10. Betrachten Sie nachfolgende Grafik:



Sind diese 2 Varianten von Entscheidungen sinngemäß identisch?

- ○ ja ○ nein

Figure C.2: Comprehension Questionnaire - Part 2

11. Welche der folgenden Aussagen treffen zu?

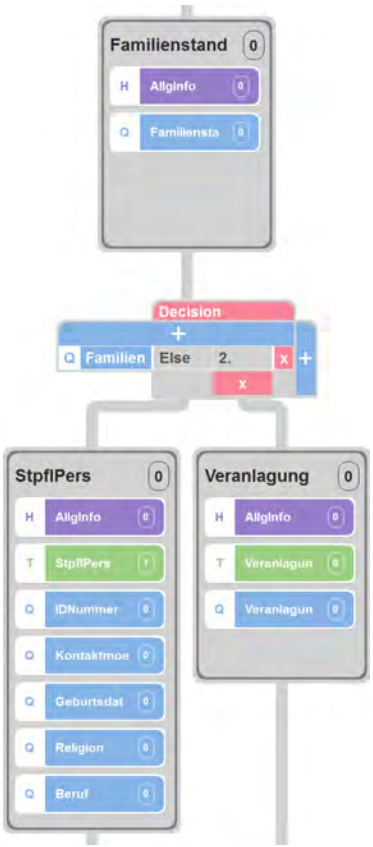
Bitte wählen Sie **einen** oder **mehrere** Punkte aus der Liste aus.

Beachten Sie, dass die Revisionsnummern bei ”0” beginnen, aber eine Version ”0” nicht existiert!

- ☐ Der Text ist in der 1ten Version
- ☐ Die Seite enthält 3 Elemente
- ☐ Die Seite ist in der 2ten Version
- ☐ 3 Elemente sind Fragen
- ☐ Das 3te Element ist ein Text
- ☐ Das 1te Element ist eine Überschrift
- ☐ Der Name der Seite ist ”Persoenlich”
- ☐ Element ”Geburtsdat” ist ein Text
- ☐ Element ”Name” ist in der 1ten Version

Figure C.3: Comprehension Questionnaire - Part 3

12. Betrachten Sie folgenden Ausschnitt eines Modells:



Welche der folgenden Aussagen treffen zu?
Bitte wählen Sie **einen** oder **mehrere** Punkte aus der Liste aus.

- ☐ Die 1te Seite enthält 4 Elemente
- ☐ Das 2te Element der 1ten Seite ist eine Frage
- ☐ Seite "Veranlagung" wird angezeigt, wenn Antwort 1 ausgewählt wurde
- ☐ Seite "Veranlagung" wird angezeigt, wenn keine Antwort ausgewählt wurde
- ☐ Seite "StpflPers" wird angezeigt, wenn nicht Antwort 2 ausgewählt wurde
- ☐ Die Entscheidung bezieht sich auf das 2te Element der 1ten Seite

Figure C.4: Comprehension Questionnaire - Part 4

D

Demographic Questionnaire

The results of the demographic questionnaire (cf. Table 3.3) are presented in Figure D.1-D.6. Familiar with QN, confident in QN and competent in QN are determined on a 7-point Likert scale ranging from fully applies (1) to does not apply at all (7).

Subject	Profession	Education Years	Gender	Course of Study
2	Apprentice/Student	17	Male	Software Engineering
3	Apprentice/Student	16	Male	Media Informatics
4	Academic	14	Female	Computer Science
5	Apprentice/Student	18	Female	Economic Sciences
6	Apprentice/Student and completed vocational training	15	Male	Computer Science
10	Apprentice/Student	19	Female	Economic Sciences
15	Apprentice/Student	18	Male	Media Informatics
18	Apprentice/Student	16	Male	Computer Science
19	Apprentice/Student	15	Male	Computer Science M. Sc.
21	Apprentice/Student	18	Male	Computer Science
24	Apprentice/Student	15	Male	Media Informatics
25	Apprentice/Student	16	Female	Media Informatics
26	Apprentice/Student	15	Male	Mathematics
28	Apprentice/Student	13	Male	Business Mathematics
29	Apprentice/Student	17	Male	Computer Science
32	Apprentice/Student	15	Female	Computer Science
33	Apprentice/Student	18	Male	Business Informatics
34	Apprentice/Student	19	Male	Computer Science M. Sc.
35	Apprentice/Student	18	Male	Media Informatics
36	Apprentice/Student	19	Male	Computer Science
37	Apprentice/Student	18	Male	Computer Science
38	Academic	17	Female	Computer Science
39	Apprentice/Student	19	Male	Media Informatics
40	Apprentice/Student	17	Male	Media Informatics

Figure D.1: Demographic Questionnaire - Novices - Part 1

Subject	Degree	Familiar With QN	Confident In QN	Competent In QN
2	University Degree	4	4	4
3	University Degree	7	7	7
4	University Degree	7	5	5
5	University Degree	7	7	5
6	Polytechnic Degree	7	4	4
10	University Degree	7	7	7
15	General Higher Education Entrance Qualification	5	6	5
18	University Degree	7	5	7
19	University Degree	7	4	6
21	General Higher Education Entrance Qualification	6	5	6
24	General Higher Education Entrance Qualification	7	7	7
25	General Higher Education Entrance Qualification	6	6	6
26	General Higher Education Entrance Qualification	7	7	7
28	General Higher Education Entrance Qualification	7	4	4
29	General Higher Education Entrance Qualification	4	4	4
32	General Higher Education Entrance Qualification	6	6	6
33	B. Sc.	7	5	6
34	University Degree	7	na	na
35	General Higher Education Entrance Qualification	7	2	4
36	General Higher Education Entrance Qualification	7	4	7
37	University Degree	7	7	7
38	University Degree	7	6	6
39	General Higher Education Entrance Qualification	6	6	4
40	General Higher Education Entrance Qualification	7	7	7

Figure D.2: Demographic Questionnaire - Novices - Part 2

Subject	Start PM (Years)	No. Process Models Analyzed/Read	No. Process Models Created/Edited	Average Activites	Amount Training Days	Self Education (Days)	Start using QN (Months)
2	2	6	0	6	0	1	0
3	3	10	4	5	10	2	0
4	4	1	0	5	0	2	0
5	1	na	na	na	na	na	0
6	2	10	5	8	0	30	0
10	1	10	0	10	30	10	0
15	3	5	2	8	20	14	0
18	2	11	4	10	1	2	0
19	2	5	0	4	2	2	0
21	2	0	0	0	0	0	0
24	0	5-7	0	7	0	0	0
25	1	3	3	20	5	6	0
26	0	0	0	na	0	0	0
28	na	na	na	na	na	na	na
29	0	0	0	na	na	na	0
32	2	2	1	na	na	10	0
33	1	2	0	na	16	4	0
34	5	<5	<5	10-30	0	0	na
35	2	18	7	8	na	na	na
36	2	0	0	0	na	0	0
37	4	8	4	10	50	20	0
38	na	na	na	na	0	0	0
39	3	15	0	20-30	0	30	0
40	2	7	3	10	0	10	0

Figure D.3: Demographic Questionnaire - Novices - Part 3

Subject	Profession	Education Years	Gender	Course of Study
1	Academic	20	Male	Business Informatics
7	Apprentice/Student	16	Male	Media Informatics
8	Apprentice/Student	18	Male	Computer Science
9	Academic	21	Male	Economic Sciences
11	Apprentice/Student	15	Male	Media Informatics
12	Apprentice/Student	14	Male	Media Informatics
13	Academic	21	Male	Computer Science
16	Apprentice/Student	19	Male	Media Informatics
17	Apprentice/Student	15	Male	Media Informatics
22	Apprentice/Student	25	Male	Computer Science
23	Apprentice/Student	15	Male	Computer Science
27	Apprentice/Student	19	Female	Economic Sciences M.Sc.
30	Academic	17	Male	Economic Sciences
31	Academic	17	Male	Software Engineering
41	Apprentice/Student	16	Male	Software Engineering M.Sc.
42	Academic	25	Male	Computer Science Dipl.
43	Academic	22	Male	Media Informatics
44	Apprentice/Student	19	Female	Computer Science M. Sc.

Figure D.4: Demographic Questionnaire - Experts - Part 1

Subject	Degree	Familiar With QN	Confident In QN	Competent In QN
1	Master of Science	7	7	7
7	University Degree	7	4	6
8	University Degree	7	4	4
9	University Degree	7	7	7
11	General Higher Education Entrance Qualification	7	7	7
12	General Higher Education Entrance Qualification	7	7	7
13	University Degree	7	4	4
16	University Degree	7	4	7
17	General Higher Education Entrance Qualification	7	7	7
22	General Higher Education Entrance Qualification	5	3	5
23	General Higher Education Entrance Qualification	7	4	4
27	University Degree	7	7	7
30	University Degree	6	5	7
31	Polytechnic Degree	7	4	4
41	Polytechnic Degree	7	7	7
42	University Degree	5	4	5
43	University Degree	7	5	7
44	B. Sc.	5	5	5

Figure D.5: Demographic Questionnaire - Experts - Part 2

Subject	Start PM (Years)	No. Process Models Analyzed/Read	No. Process Models Created/Edited	Average Activites	Amount Training Days	Self Education (Days)	Start using QN (Months)
1	10	100	10	25	0	2	0
7	2	20	10	15	15	50	0
8	1	30	20	20	na	na	0
9	1	15	10	<20	15	15	0
11	1	35	13	15	2	3	0
12	0.5	40	20	17	0	10	0
13	4	>100	50	20	10	5	0
16	2	30	15	11	15	15	0
17	1	40	20	10	3	40	0
22	2	20	10	10	0	5	0
23	1	15	10	15	5	5	0
27	1	30	15	10	150	50	0
30	1	20	10	8	60	60	1
31	na	10	10	5	2	4	na
41	4	20-30	15-20	25	30	20	na
42	5	10	10	10	0	20	0
43	6	100	30	15	na	100	0
44	6	20	15-20	10-20	25	15	na

Figure D.6: Demographic Questionnaire - Experts - Part 3

E

Raw Data

Figures E.1-E.4 present detailed evaluation results from each questionnaire model. ME Model, ME Change Request, ME Correctness (referred to Table 3.4) are determined on a 7-point Likert scale ranging from fully applies (1) to does not apply at all (7). The use of screencasts is presented as ME Screencasts (1 = yes, 2 = no).

Subject	No. Activities	Total Time	No. Errors	CR Time	ME Model	ME Change Request	ME Correctness	ME Screncasts
2	12	512.25	0	286.241	6	5	2	1
3	12	742.777	2	200.482	5	6	2	2
4	14	500.528	1	97.793	5	7	2	2
5	8	523.314	0	188.152	6	6	3	2
6	8	584.845	4	361.972	5	4	1	2
10	6	504.071	5	na	2	1	6	1
15	10	870.523	4	449.948	2	2	5	1
18	10	160.307	1	87.663	6	6	1	2
19	12	778.407	4	301.99	2	3	1	1
21	12	446	2	80.746	5	6	2	2
24	8	541.807	4	347.643	6	5	2	2
25	8	561.627	4	384.321	3	2	2	2
26	8	469.522	6	383.301	6	4	3	2
28	7	572.186	9	358.037	6	6	1	2
29	14	431.828	0	170.621	5	7	6	1
32	8	533.908	4	191.195	5	3	4	1
33	8	294.857	4	178.801	6	7	2	2
34	12	691.453	0	405.292	4	4	2	1
35	8	264.781	4	175.983	5	6	1	2
36	8	553.724	4	430.914	3	3	2	2
37	8	341.659	4	224.039	6	5	2	2
38	7	1073.568	10	806.418	3	2	2	1
39	10	522.243	1	274.405	6	5	2	2
40	10	563.792	0	132.845	5	7	1	2

Figure E.1: Raw Data - Novices - Task 1

Subject	No. Activities	Total Time	No. Errors	CR Time	ME Model	ME Change Request	ME Correctness	ME Screncasts
2	17	326.225	1	17.63	6	5	2	1
3	13	764.808	0	483.679	5	3	1	2
4	13	335.667	0	66.628	7	7	2	2
5	19	659.953	3	190.516	6	6	7	2
6	13	530.919	0	192.166	3	5	1	2
10	11	1197.167	11	541.26	1	1	7	1
15	18	1311.955	2	443.76	3	3	5	2
18	13	172.94	2	66.165	4	5	2	2
19	15	634.628	5	247.295	4	4	4	2
21	17	861.97	0	495.749	3	3	2	1
24	11	489.731	1	324.809	5	3	2	2
25	13	509.929	10	135.463	3	3	3	2
26	22	322.907	1	94.247	5	7	2	2
28	31	1441.049	7	1150.76	2	2	2	2
29	11	709.107	0	353.299	4	6	6	1
32	15	628.193	2	367.945	3	2	5	2
33	26	823.903	2	336.041	6	6	2	2
34	21	1183.565	1	547.671	4	4	2	1
35	9	148.218	1	22.802	6	6	1	2
36	8	613.756	15	411.529	2	2	5	2
37	17	363.087	1	152.393	6	5	2	2
38	13	645.312	12	241.991	4	3	3	1
39	15	516.331	0	172.099	3	3	3	2
40	13	537.695	0	336.962	4	5	1	2

Figure E.2: Raw Data - Novices - Task 2

Subject	No. Activities	Total Time	No. Errors	CR Time	ME Model	ME Change Request	ME Correctness	ME Screncasts
1	14	761.787	1	154.92	2	5	2	2
7	10	409.365	3	195.81	5	3	2	2
8	10	758.253	1	160.933	3	6	2	2
9	10	1023.751	4	648.326	4	3	2	1
11	12	905.533	1	341.524	5	4	2	2
12	10	733.771	0	155.438	6	5	2	2
13	12	521.196	0	105.795	3	6	2	2
16	10	749.571	0	132.041	5	6	1	2
17	8	425.244	4	266.738	6	5	2	2
22	8	468.848	4	368.858	6	5	2	1
23	10	529.685	1	232.828	4	4	2	2
27	16	864.346	0	452.027	3	2	5	1
30	6	794.066	2	na	3	6	3	1
31	16	489.838	0	197.218	6	6	1	2
41	10	646.805	0	226.704	3	3	4	2
42	8	337.863	4	145.152	3	4	3	2
43	10	153.215	6	na	5	3	2	2
44	9	555.347	6	na	4	2	5	1

Figure E.3: Raw Data - Experts - Task 1

Subject	No. Activities	Total Time	No. Errors	CR Time	ME Model	ME Change Request	ME Correctness	ME Screncasts
1	15	797.376	5	174.525	3	5	2	1
7	9	516.719	0	164.545	4	5	3	1
8	9	482.488	0	255.379	4	3	2	2
9	13	816.637	1	488.033	3	2	4	1
11	9	778.46	1	463.634	4	2	3	2
12	15	761.153	0	240.844	4	3	2	1
13	49	970.835	0	240.392	2	3	2	2
16	10	620.967	7	356.834	5	5	1	2
17	13	750.97	0	381.028	5	5	3	2
22	17	466.23	1	130.728	6	4	2	2
23	11	536.808	1	326.842	5	5	3	2
27	21	630.991	9	161.41	3	2	7	1
30	12	760.172	0	339.495	3	2	3	1
31	15	358.491	7	81.96	6	6	3	2
41	15	554.078	0	289.229	4	4	3	2
42	13	414.483	1	231.317	5	5	3	2
43	23	638.713	8	107.854	3	2	2	2
44	7	444.433	10	310.503	4	4	2	2

Figure E.4: Raw Data - Experts - Task 2

F

Test for Normal Distribution

The following figures and tables present the distribution of the data. In particular, Figure F.1 graphically presents the results for the total time regarding the group of novices while Table F.1 shows the p-values of two different statistical tests of normality. In the same way, the results for the total time regarding the group of experts are presented in Figure F.2 as well as in Table F.2. The results regarding the number of errors are presented in Figure F.3 and Table F.3 (novices) as well as in Figure F.4 and Table F.4 (experts).

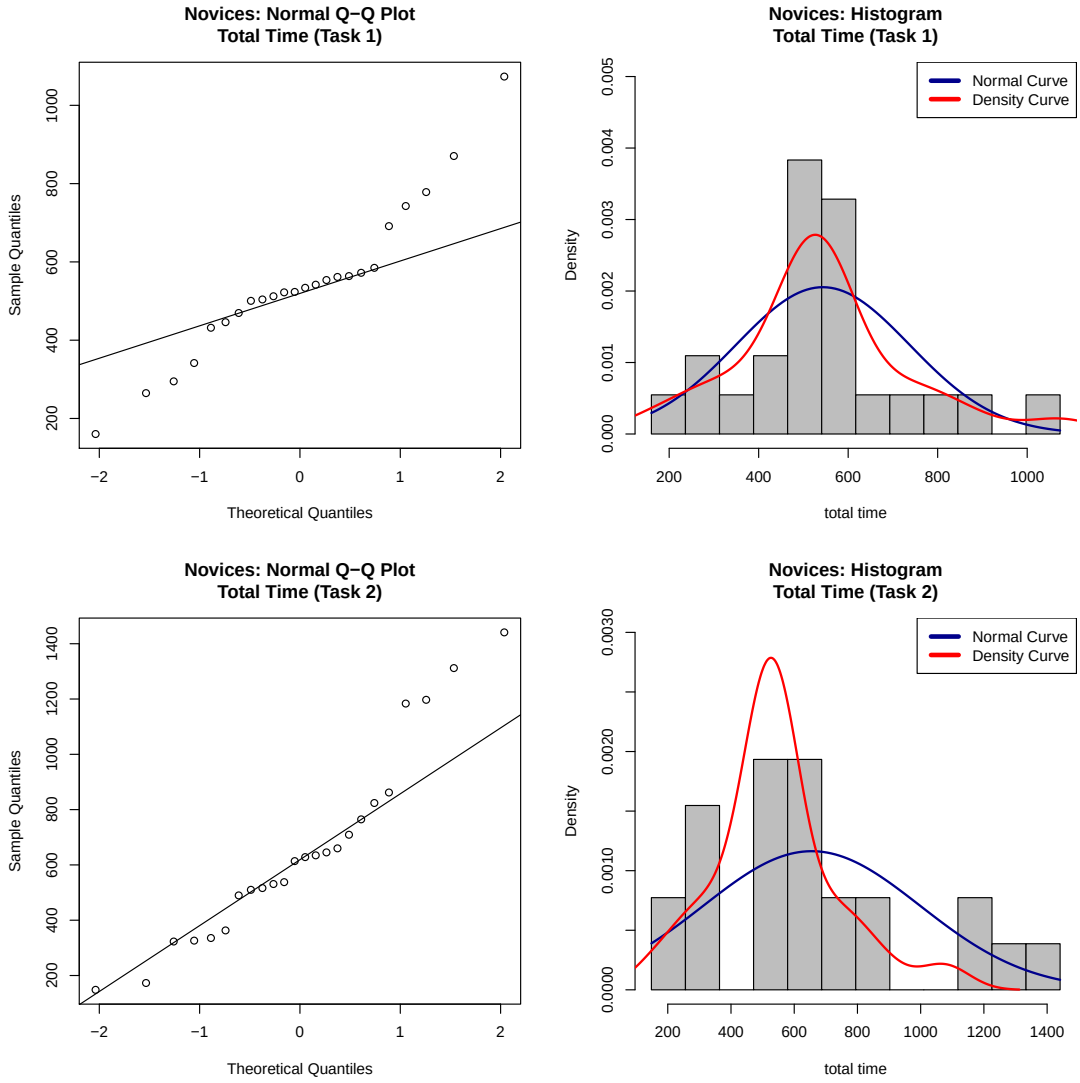


Figure F.1: Novices: Distribution of Total Time

Task	Shapiro–Wilk test	Anderson–Darling test
Task 1	0.1415242	0.0452340
Task 2	0.0735861	0.0572161

Table F.1: Novices: P-Values Total Time

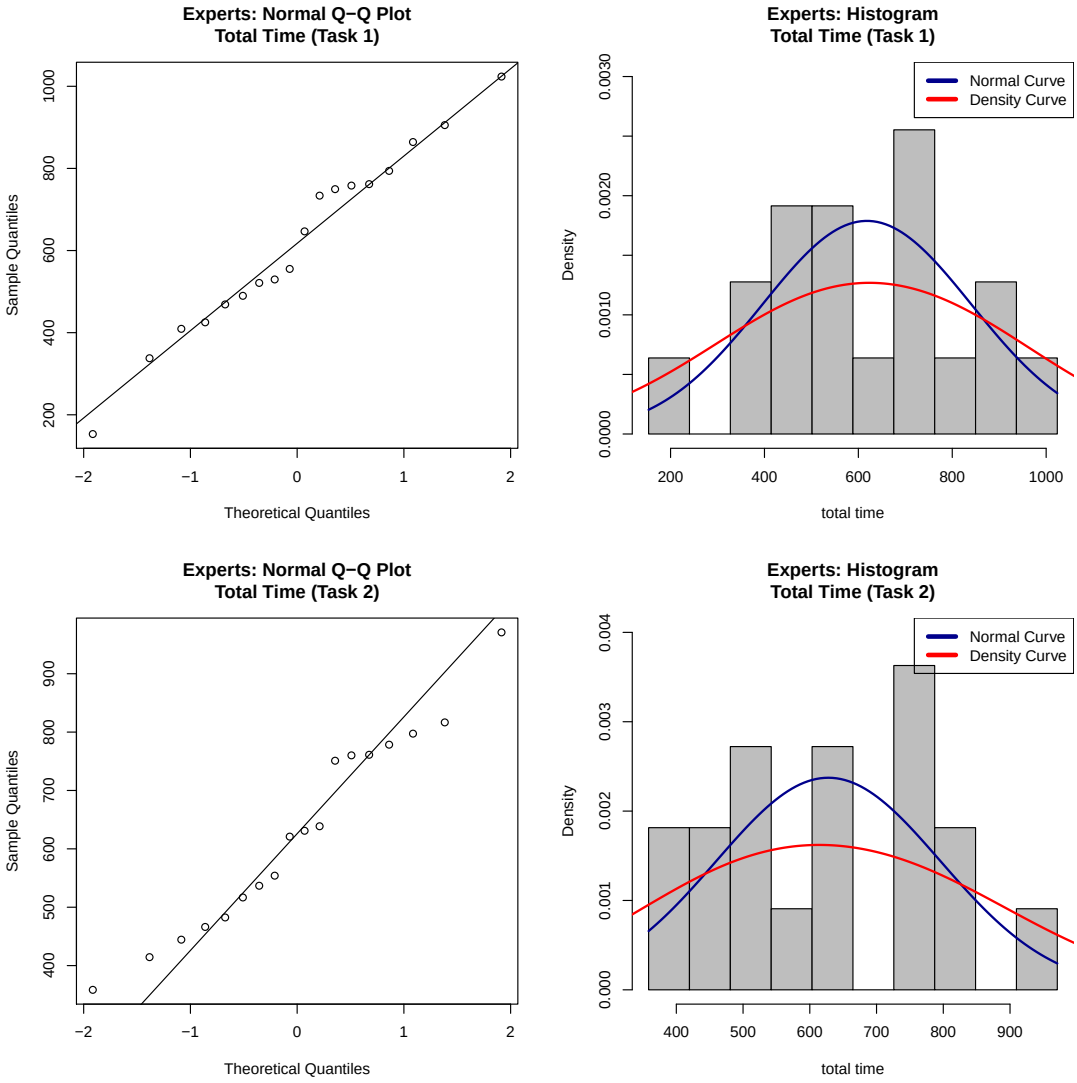


Figure F.2: Experts: Distribution of Total Time

Task	Shapiro–Wilk test	Anderson–Darling test
Task 1	0.945432	0.765591
Task 2	0.598748	0.461717

Table F.2: Experts: P-Values Total Time

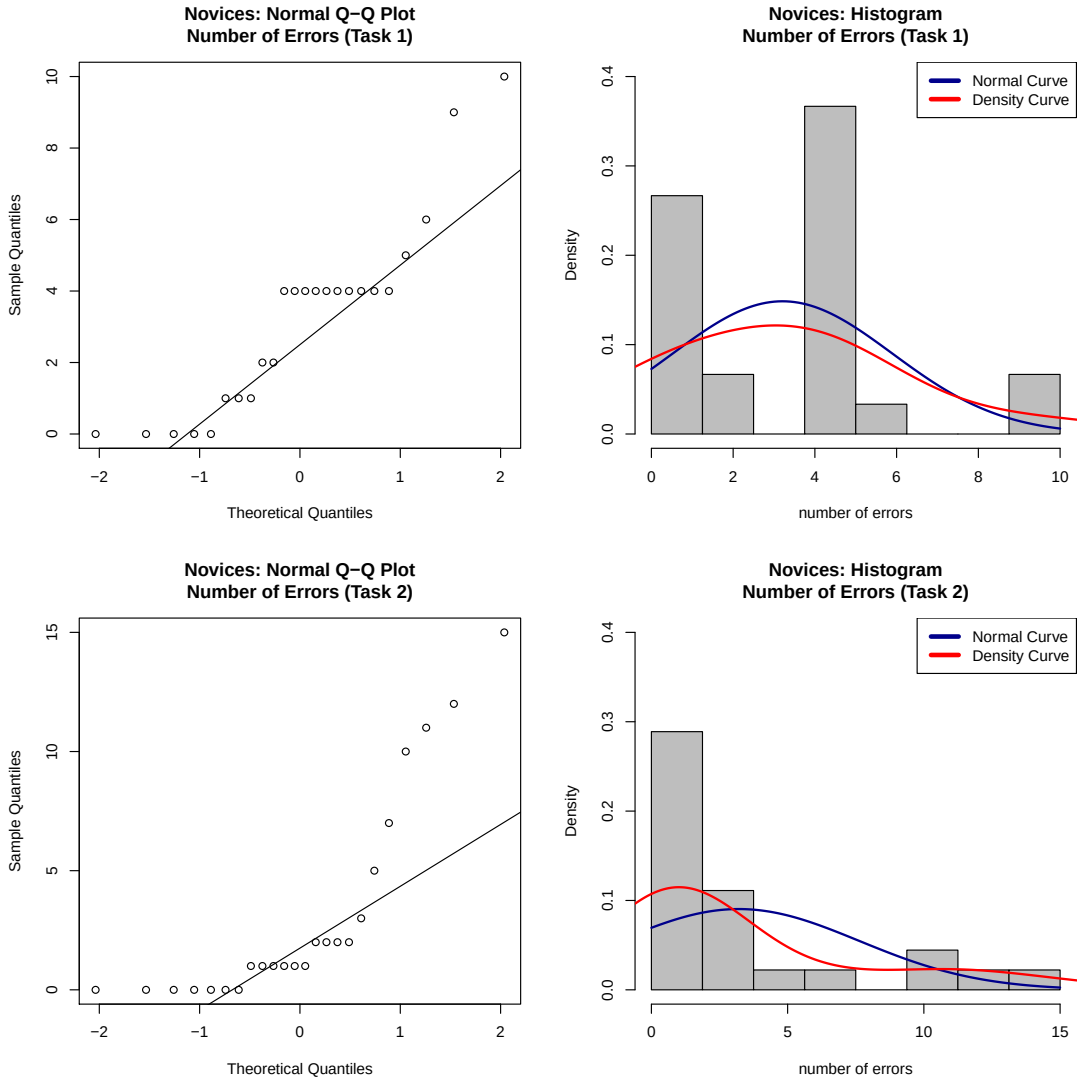


Figure F.3: Novices: Distribution of Number of Errors

Task	Shapiro–Wilk test	Anderson–Darling test
Task 1	0.0048353	0.003329905
Task 2	0.0000251	0.000000402

Table F.3: Novices: P-Values Number of Errors

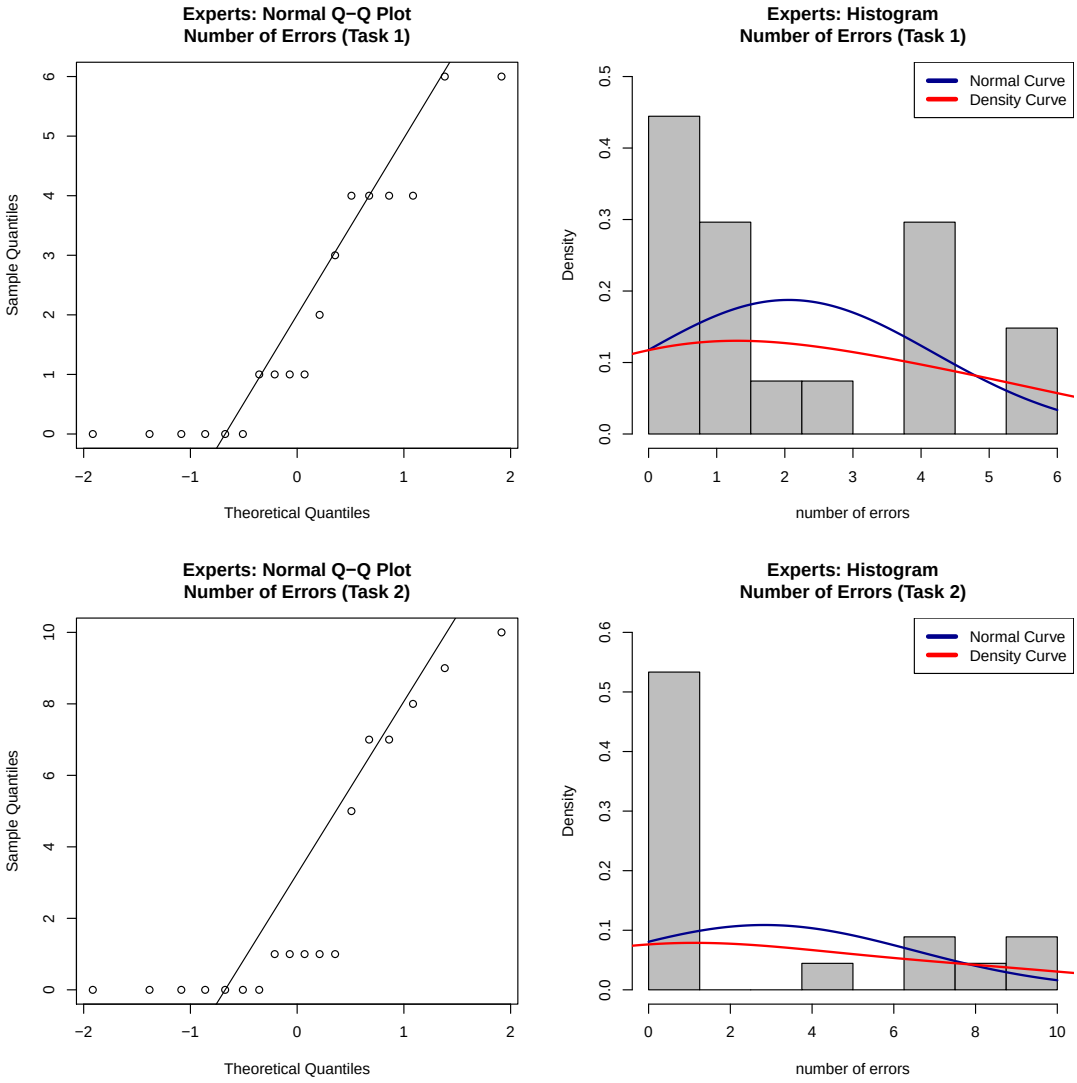


Figure F.4: Experts: Distribution of Number of Errors

Task	Shapiro–Wilk test	Anderson–Darling test
Task 1	0.006242	0.0052271
Task 2	0.000299	0.0000175

Table F.4: Experts: P-Values Number of Errors

G

Additional Experimental Results

Figures G.1-G.4 illustrate the dependencies between additional variables calculated as described in Section 5.1. In particular, Figure G.1 and Figure G.2 show the dependencies between the results regarding the group of novices separately for each task while Figure G.3 and Figure G.4 present the results in the same way for the group of experts.

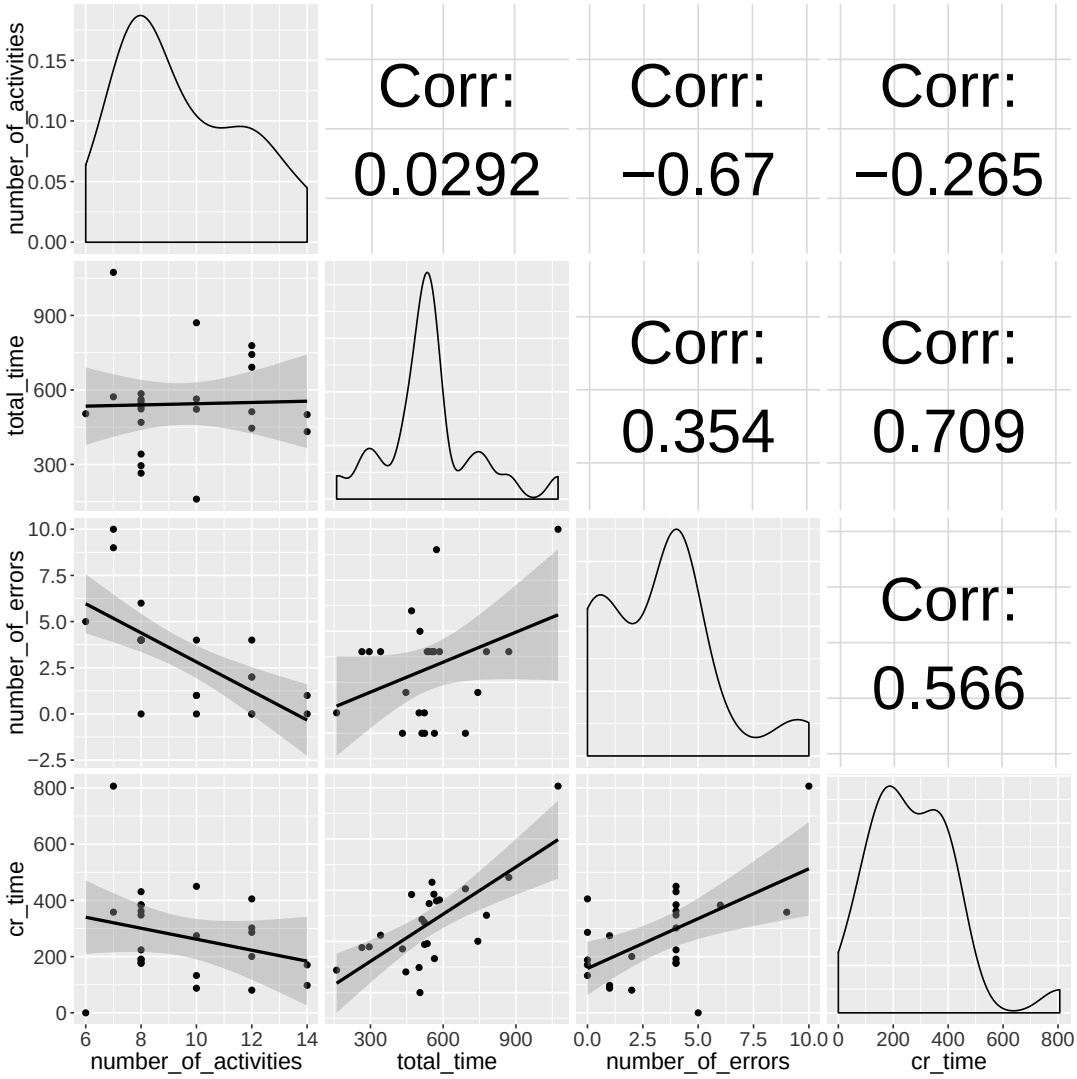


Figure G.1: Matrix Plot Novices Task 1

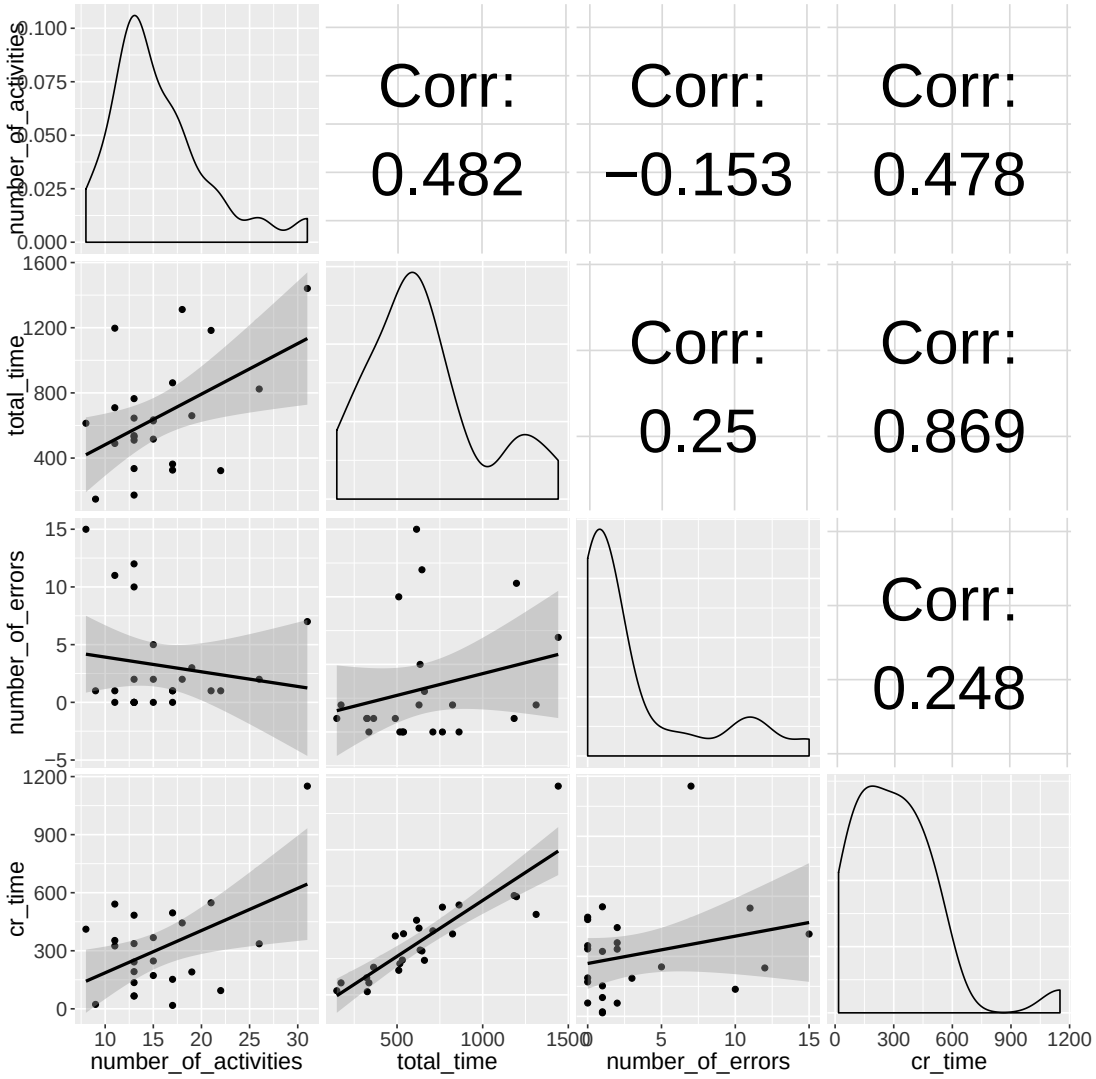


Figure G.2: Matrix Plot Novices Task 2

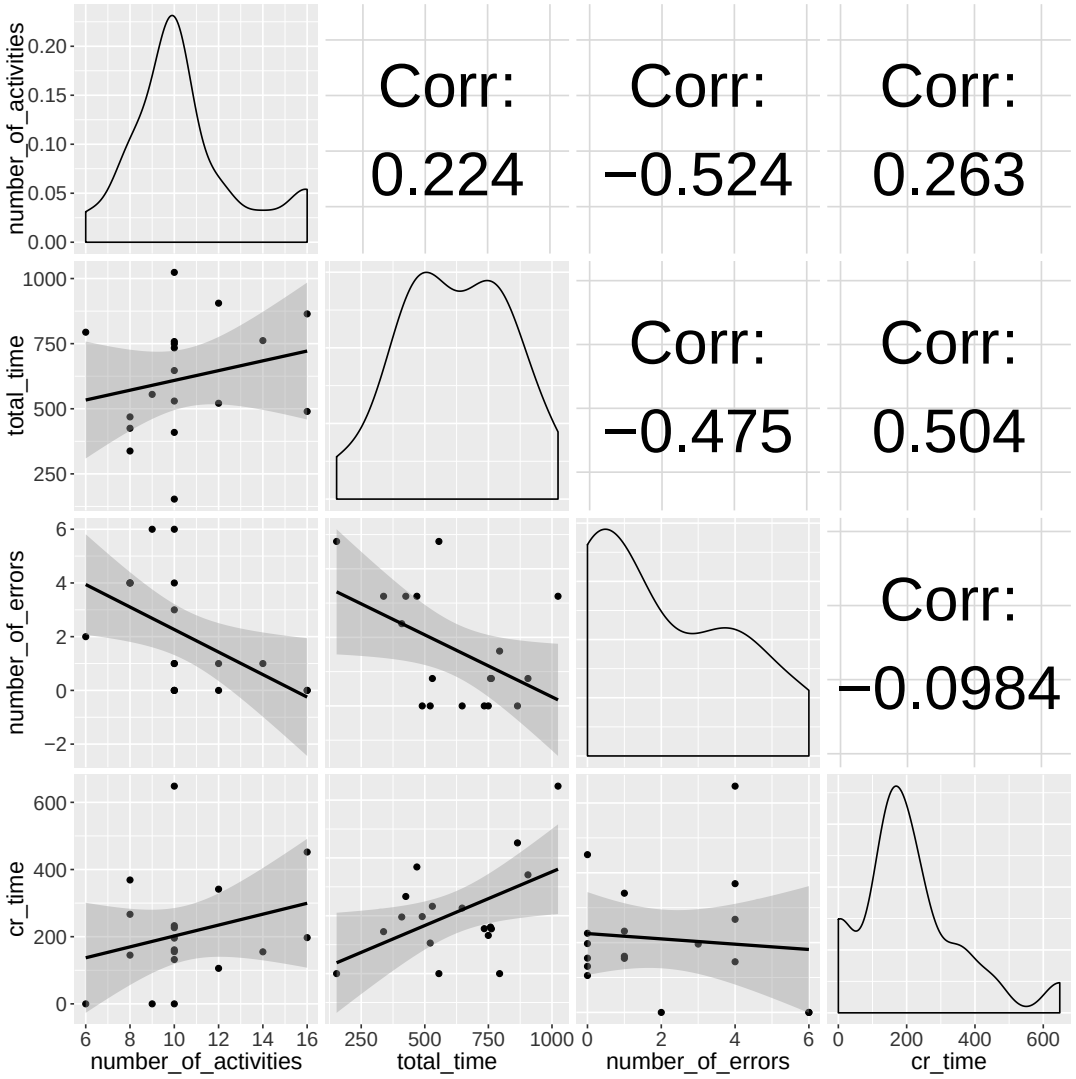


Figure G.3: Matrix Plot Experts Task 1

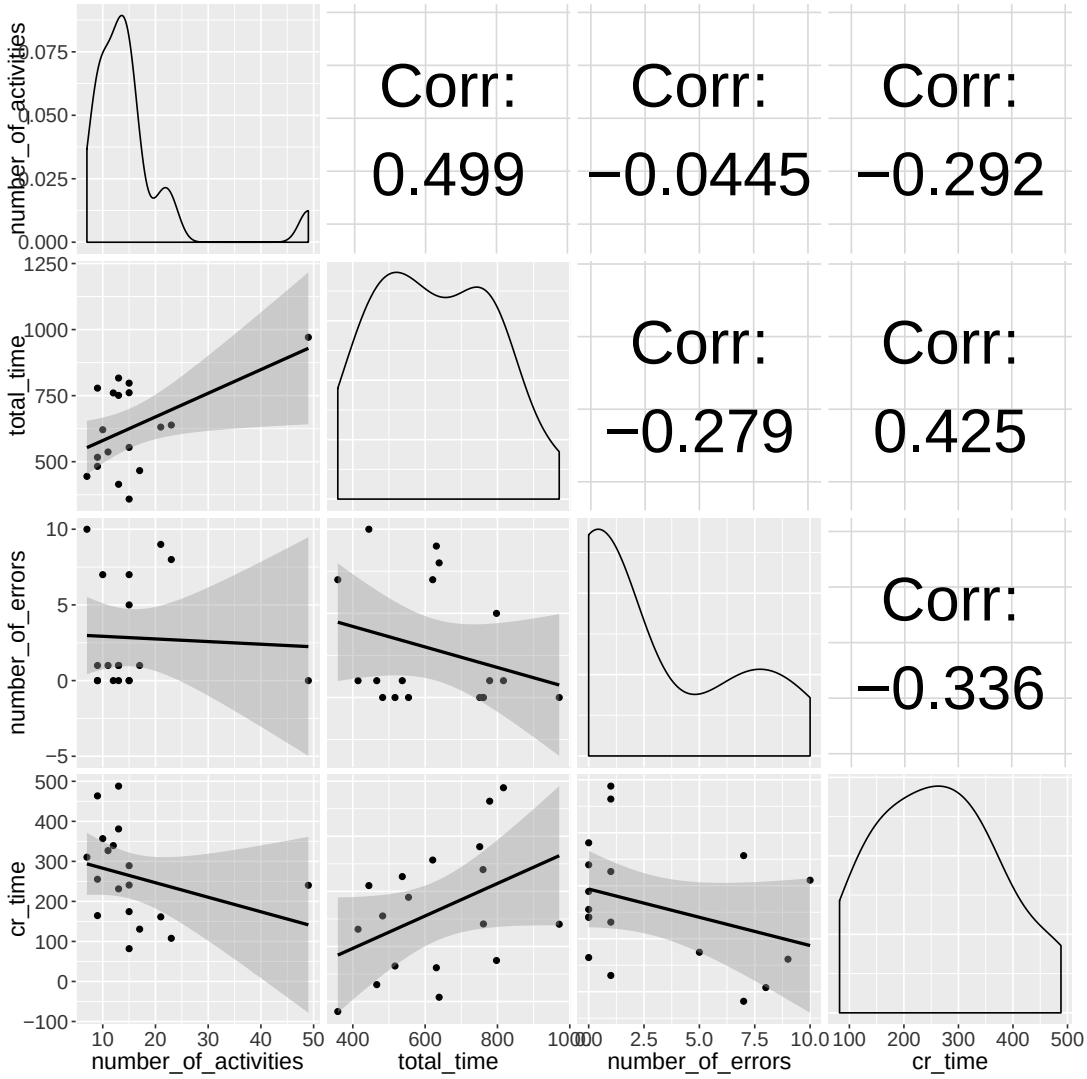


Figure G.4: Matrix Plot Experts Task 2

List of Figures

List of Tables

Name: Dominic Gebhardt Matrikelnummer: 601652

Erklärung

Ich erkläre, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Ulm, den .....

Dominic Gebhardt